7-30-2013

Handling Data with Three Types of Missing Values

Jennifer Boyko
University of Connecticut, jen.boyko@gmail.com

Handling Data with Three Types of Missing Values

Jennifer A. Boyko, Ph.D.
University of Connecticut, 2013

ABSTRACT

Missing values present challenges in the analysis of data across many areas of research. Handling incomplete data incorrectly can lead to bias, over-confident intervals, and inaccurate inferences. One principled method of handling incomplete data is multiple imputation. This dissertation considers incomplete data in which values are missing for three qualitatively different reasons and applies a modified multiple imputation framework in the analysis of that data. The first major contribution of this dissertation is a derivation of the methodology for implementing multiple imputation in three stages. Also included is a discussion of extensions to estimating rates of missing information and ignorability in the presence of three types of missing values. Simulation studies accompany these sections to assess the performance of multiple imputation in three stages. Finally, this new methodology is applied to an insomnia treatment study with comparisons to other commonly used missing data methods.

Handling Data with Three Types of Missing Values

Jennifer A. Boyko

B.S., Mathematics, Haverford College, PA, USA, 2009

M.S., Statistics, University of Connecticut, CT, USA, 2012

A Dissertation
Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy
at the
University of Connecticut

2013

APPROVAL PAGE

Doctor of Philosophy Dissertation

Handling Data with Three Types of Missing Values

Presented by

Jennifer A. Boyko, B.S. Mathematics, M.S. Statistics

Major Advisor ____________________________________

Ofer Harel

Associate Advisor ____________________________________

Nitis Mukhopadhyay

Associate Advisor ____________________________________

Sangwook Kang

University of Connecticut

2013

Dedication

For my dad, Robert J. Boyko, who would have been so proud.

Acknowledgements

Writing a dissertation is an activity which is far from solitary and could never have been accomplished without the help and support of the many wonderful people in my life. Each and every person acknowledged here contributed immeasurably to this journey, both academically and emotionally. The words here will not adequately capture my gratitude but, most simply, you have my thanks.

I came to UConn with the intention of working with Ofer Harel and I consider it one of the better decisions made in my life. I could not have asked for an advisor so in tune with the big picture of advancing my career academically and professionally. His advice, advocacy for his students, and inherent knowledge of how to best utilize my skills have been indispensable throughout my years here. It may have taken some time, but one of my greatest graduate school lessons was learning to say "Yes, Ofer. You were right."

Among the professors at UConn I must thank are Sangwook Kang, for being an associate advisor and offering comments on my dissertation, and Nitis Mukhopadhyay, for not only being an associate advisor but also for being a true inspiration. Dr. Mukhopadhyay is a challenging professor, a brilliant statistician, and a genuinely kind individual who has my utmost respect and admiration. It has been my privilege to get to know him and hear some of his stories. I am also grateful to the secretaries, Cathy Brown, Tracy Burke, and Megan Petsa, because the department could not run without them.

I would never have discovered statistics without the guidance of my Haverford professors and REU advisors. Lynne Butler, who knew I was destined to be a statistician before I even knew what statistics was, challenged me and found opportunities for me that truly steered the course of my future. Joshua Sabloff, who tried so hard to teach me analysis, offered terrific advice about being a graduate student and reminded me how important it is to keep having fun, even when writing a dissertation. My academic turning point was a summer REU program with Scott Linder. He has my eternal gratitude for taking a chance on a student who had no experience with statistics and for instilling a love of the subject. My first exposure to missing data and multiple imputation was with Jayson Wilbur and the WPI REU program. Five years and a dissertation later, it is still my favorite topic in statistics.

Throughout my graduate career, I have learned just as much from my colleagues as I have from my professors. First, I must thank two of my fellow Ofer minions: Valerie Pare and Ashok Chaurasia. Val is practically a co-author on this dissertation for all of the help she has given me. Together we coded, discussed, edited, commiserated, laughed, studied, clustered, just imputed, wrote exams, and took long lunches and coffee breaks. We shared our triumphs, frustrations, confusions, and stupid questions and she is truly my departmental sister. There is no other way to say it: I never would have survived any of this without her. Ashok offered indispensable help with coding and big ideas. As the more theoretical comrade-in-arms, he assisted me where my knowledge and experience were sorely lacking. I also learned from the successes, challenges, and celebrations of the “old people”, especially Gregory Matthews, Tony Jiang, Patrick Harrington, Shan Hu, Marcos Prates, and Sankha Perera. Equally valuable was the contribution of the “young people”, especially Gregory Vaughan, Brian Bader and Paul McLaughlin, who reminded me how far I have come and how much I have learned in my time at UConn.

The support of my friends means more to me than I can express. Mary Mulqueen, who was by my side during college and the graduate school application process, continued to boost my spirits through these four years with her encouragement and visits. Laura Pontrella, my friend for over twenty years, offered necessary respites from the world of mathematics with shared dinners, pedicures, and episodes of Revenge. With love and laughter, I acknowledge three friends who have supported me and believed in me ever since high school: Nicholas Principe, Emily Cibelli, and Devin Toohey. I would feel lost without their friendship and some of my fondest memories include commiserating over graduate school together, learning how to talk to each other about our varying fields, and celebrating the achievements of our lives with Christmas parties, cocktails, and reminiscing. I have always appreciated the advice that a dissertation should probably be started before the day it is due and that they were some of the first to refer to me as “the future Dr. Jen.”

Finally, I must thank the three most important people in my life. Zachary McGinniss, the love of my life, has been my anchor to reality. Zak supported me, made sure I remembered to eat, knew when a day had been bad enough for a necessary glass of wine, and, most of all, gave me regular pep talks reminding me that I could do this and

that earning my doctorate would be worthwhile in the end. I am blessed to be spending the rest of my life with him. My sister, Christine, has always been there for me with love, dancing, and Disney songs. She has grown up to be my best friend and I am lucky to have the most amazing sister.

I could not have gotten to this point without my first and best role model, the first Dr. J. Boyko, my mom. She inspires me daily and if I become half the woman she is, personally or professionally, I will consider myself a success. My mom is everything I am and everything I hope to be. I have spent months trying to find the right words to express my appreciation but it turns out that the sentiment can be expressed in the most simple and heartfelt manner: thank you and I love you.

Contents

List of Tables

List of Figures

Chapter 1

Introduction and Literature Review

Incomplete data is a common obstacle to the analysis of data in a variety of fields. Missing values can occur for several different reasons including failure to answer a survey question, dropout, planned missing values, intermittent missed measurements, latent variables, and equipment malfunction. In fact, many studies will have more than just one type of missing value. Appropriately handling missing values is critical in the inference for a parameter of interest. Many methods of handling missing values inappropriately fail to account for the uncertainty due to missing values. This failure to account for uncertainty can lead to biased estimates and over-confident inferences.

Multiple imputation is one method for handling incomplete data that accounts for the variability of the incomplete data. This procedure does so by filling in plausible values several times to create several complete data sets and then appropriately combining complete data estimates using specific combining rules. This method is praised for the ability to use complete data analytical methods on each data set as well as for retaining the variability seen in the observed data to arrive at estimates which are not distorted by the imputation method.

Often there is more than one type of missing value in a study. Typically, all of the missing values are treated as though they are the same type. However, there are benefits to treating each of these types of missing values separately. One benefit arises when imputing one type of missing value first computationally simplifies the imputation of the rest of the missing values. A second important benefit is that treating the two types differently can allow the researcher to quantify how much variability and how much missing information is due to each type of missing value. A third benefit is the ability to assign different assumptions to each of the missing values. That is, each of the missing value types might have different assumptions on the mechanisms generating the missingness. Two-stage multiple imputation is a nested version of multiple imputation where missing values of one type are imputed first. For each imputation of the first type of missing value, additional imputations are created for the second type of missing value, holding the first imputed values fixed. Separate combining rules determine how estimates should be handled to lead to valid inferences.

One area which is still unexplored is the situation where there are three types of missing values in a study. Studies which tend to have large amounts of missing values also tend to have missing values of several different types. Development of a three-stage multiple imputation approach would be beneficial in analyzing these types of studies. Three-stage multiple imputation would also extend the benefits of two-stage multiple imputation, namely the quantification of the variability attributable to each type of missing value and the flexibility for greater specificity regarding data analysis.

1.1 Overview of Missing Data

Missing data refers to unobserved values in a data set which can be of different types and may be missing for different reasons. These different reasons can include unit nonresponse, item nonresponse, dropout, human error, equipment failure, and latent classes. Missing data can be problematic for researchers across many different fields. Measures should always be taken to avoid incidence of missing data but appropriate missing data methods typically need to be considered when analyzing incomplete data (Harel et al., 2012). Knowledge about the nature of the missing values can help identify the most appropriate method for dealing with missing data (Little & Rubin, 2002).

1.1.1 Pattern and Mechanism of Missingness

In order to characterize the missingness, the pattern and mechanism of missingness are used. The pattern of missingness visually illustrates how values are missing in a data set while the mechanism of missingness deals with the probabilistic definition of the missing values.

Pattern of Missingness

The pattern of missingness maps which values are missing in a data set to assist in identifying patterns and interdependencies in the incomplete data. Some of the patterns of missingness described by Schafer & Graham (2002) include univariate, monotone, and arbitrary patterns. A univariate pattern would refer to the situation in which the missing values occur in only one of the variables while the other variables are completely observed. The monotone pattern is a dropout pattern. That is, if a subject has a missing value in the i^{th} position, then all of the subsequent values are also missing. The arbitrary pattern of missingness is exactly as the name implies: the missing values occur in any of the variables in any position.

Mechanism of Missingness

The mechanism of missingness describes the relationship between the probability of a value being missing and the other variables in the data set.

Let Y represent the complete data that can be partitioned as (Y_{obs}, Y_{mis}) where Y_{obs} is the observed part of Y and Y_{mis} is the missing part of Y. Let R be an indicator random variable (or matrix) indicating whether or not Y is observed or missing. Let $R = 1$ denote a value which is observed and let $R = 0$ denote a value which is missing. The statistical model for missing data is $P(R|Y, \phi)$ where ϕ is the parameter for the missing data process. The mechanism of missingness is determined by the dependency of R on the variables in the data set. The following are the mechanisms of missingness as described in Rubin (1976) and Rubin (1987).

The first mechanism of missingness is *missing at random* (MAR). This mechanism of missingness is given by

$$P(R|Y, \phi) = P(R|Y_{obs}, \phi).$$

That is, the probability of missingness is only dependent on observed values in Y and not on any unobserved values in Y. A simple example of MAR is a survey where subjects over a certain age refuse to answer a particular survey question and age is an observed covariate.

The second mechanism of missingness is a special case of MAR known as *missing completely at random* (MCAR). In this case, the mechanism of missingness is given by

$$P(R|Y, \phi) = P(R, \phi).$$

The probability of missingness is not dependent on any observed or unobserved values in Y. It is what one colloquially thinks of as "random." One example of MCAR might be a computer malfunction that arbitrarily deletes some of the data values.

The third mechanism of missingness is referred to as *missing not at random* (MNAR). This mechanism of missingness occurs when the conditions of MAR are violated so that the probability of missingness is dependent on Y_{mis} or some unobserved covariate. One instance of MNAR might be subjects who have an income above a certain value refusing to report an income in the survey. Here the missingness is dependent on the unobserved response, income.

Ignorable and Nonignorable Nonresponse Mechanisms

A missing data mechanism can also be classified as ignorable or nonignorable. In order for a missing data mechanism to be ignorable, two conditions must be met. The first condition is that the data are MAR or MCAR. The second condition is that the parameter of interest, θ, and the parameter of the missing data process, ϕ, are distinct. These parameters are considered distinct if the following condition holds:

$$P(\theta, \phi) = P(\theta)P(\phi),$$

that is, the joint prior distribution for θ and ϕ is equal to the product of two independent priors. From a frequentist perspective, the parameters are distinct if the joint parameter space is the Cartesian cross-product of the individual parameter spaces of θ and ϕ (Schafer, 1997). A missing data mechanism is classified as nonignorable if at least one of these conditions is not met. Ignorability represents the weakest set of conditions under which the distribution of R, the missingness, does not need to be considered in Bayesian or likelihood-based inference of θ (Rubin, 1987).

1.1.2 Methods for Handling Incomplete Data

Early Methods

Case deletion, or complete case (CC) analysis is a method of dealing with missing values that is commonly used as the default in many statistical packages. In this approach, only completed cases with no missing values are included in the analysis. Several papers (Harel et al., 2012; White & Carlin, 2010; Belin, 2009) show examples of using CC and producing biased results with low power. Additionally, it is costly to obtain data and to simply omit partially completed data is wasteful and results in a loss of information and a loss of money. CC is only an appropriate method to use when the missing values are a truly random subset of the complete data.

A second missing data method is single imputation. Instead of using CC, researchers seek to impute, or fill in, missing values with plausible values. There exist a wide range of single imputation values which fill in one value for each missing value. These methods are ad hoc approaches which can range from plausible to destructive. One method of single imputation is imputing unconditional means which simply means that missing values are replaced with the average of the observed values. However, this approach drastically reduces variance. A second approach is hot deck imputation which replaces missing values with random draws from the observed data. This approach runs the risk of distorting correlations and measures of association. A third approach is conditional mean imputation. In this case, missing values are replaced with fitted values from a least squares regression line. The problem with this approach is in overstating the relationship between the independent and dependent variables. Another approach commonly used in longitudinal studies is last observation carried forward (LOCF). This method replaces all missing values with the last observed value for each individual subject. LOCF can be plausible for some studies but can cause gross bias in studies where natural decline occurs. In those cases, LOCF can imply that a subject is failing to decline at a natural rate which is obviously unrealistic (Molnar et al., 2008).

These ad hoc methods are not solidly grounded in mathematical foundations and exist merely for their ease of implementation. Generally speaking, single imputation causes a reduction in the natural variance of the data that can distort inferences. Many of these single imputation methods are thoroughly described in Schafer & Graham (2002).

Alternative Methods

One alternative to the ad hoc techniques is multiple imputation (Rubin, 1976). Multiple imputation is a common method of dealing with missing data which involves creating several complete data sets and appropriately combining parameter estimates and variances. Multiple imputation is one of the main focuses of this dissertation and is discussed in depth in Section 1.2.

Other sophisticated methods of dealing with missing values include weighting techniques (Meng, 1994), maximum likelihood (Little & Rubin, 2002) via the EM algorithm (Dempster et al., 1997), and Bayesian analysis (Gelman et al., 2003).

1.2 Multiple Imputation

1.2.1 Standard Multiple Imputation

The idea behind multiple imputation is to fill in plausible values for the missing data several times to account for model uncertainty (Rubin, 1987; Harel & Zhou, 2007). After creating m complete data sets by drawing from the posterior predictive distribution of the missing values, each data set is analyzed using complete data analysis methods. Let Q denote the parameter of interest. An example of such a Q might be a mean or a regression coefficient. From the complete data analyses, complete data estimates ($\hat{Q}$) and their associated variances (U) are obtained.

Combining Rules

Let $Y = (Y_{obs}, Y_{mis})$ be the complete data where Y_{obs} is the observed part of the data and Y_{mis} is the missing part of the data. The actual posterior distribution of Q can be represented as the complete data posterior distribution of Q averaged over the posterior distribution of the missing data (Rubin, 1987) as follows:

$$P(Q|Y_{obs}) = \int P(Q|Y_{obs}, Y_{mis}) P(Y_{mis}|Y_{obs}) dY_{mis}.$$

The consequences that follow lead to the combining rules of multiple imputation. The first is regarding the final estimate of Q where

$$E(Q|Y_{obs}) = E[E(Q|Y_{obs}, Y_{mis})|Y_{obs}]$$

meaning that the posterior mean of Q is equal to the average of the repeated complete data posterior means of Q. The next consequence is regarding the posterior variance of Q being the sum of the average of the repeated imputation variances and the variance of the repeated imputation posterior means of Q (Rubin, 1987). Mathematically speaking, this is

$$V(Q|Y_{obs}) = E[V(Q|Y_{obs}, Y_{mis})|Y_{obs}] + V[E(Q|Y_{obs}, Y_{mis})|Y_{obs}]. \tag{1.1}$$

The original derivations for the combining rules were based on large sample inference (Rubin, 1987). The paper by Reiter & Raghunathan (2007) review the implications of basing the derivation on large sample inference. The assumption involved was that, in the presence of the complete data, intervals and tests would be based on a normal approximation. That is,

$$(\hat{Q} - Q)/\sqrt{U} \sim N(0, 1).$$

The overall estimate of Q is

$$\bar{Q} = m^{-1} \sum \hat{Q}^{(j)}$$

where $\hat{Q}^{(j)}$ is the estimate from the j^{th} repeated imputation. To get the standard error for $\bar{Q}$, the between-imputation variance and the within-imputation variance must be appropriately combined. The between-imputation variance is denoted by B and is

$$B = (m-1)^{-1} \sum (\hat{Q}^{(j)} - \bar{Q})^2$$

while the within-imputation variance is denoted by $\bar{U}$ and is

$$\bar{U} = m^{-1} \sum U^{(j)}$$

where $U^{(j)}$ is the estimated variance of $\hat{Q}^{(j)}$.

The total variance, denoted by T is then equal to

$$T = \bar{U} + (1 + m^{-1})B. \tag{1.2}$$

In equation (1.2), the $(1 + m^{-1})B$ estimates the increase in variance because of the missing data and $\bar{U}$ estimates the variance if the data were complete (Reiter & Raghunathan, 2007).

In an ideal situation, where there could be an infinite number of imputations,

$$(\bar{Q} - Q)/\sqrt{T} \tag{1.3}$$

would have a $N(0,1)$ distribution. However, to account for the finite number of imputations and for the simulation error associated with that finite number of repetitions, equation (1.3) follows a t_ν distribution. The degrees of freedom associated with this distribution is found by matching the first two moments of a chi-squared distribution to yield

$$\nu = (m-1)\left(1 + \frac{\bar{U}}{(1+m^{-1})B}\right)^2 \tag{1.4}$$

(Rubin, 1987; Schafer, 1999). It is worth noting that if Y_{mis} carries no information about Q, then T is reduced to $\bar{U}$.

There is additional work that has been done regarding the degrees of freedom as represented in equation (1.4). The risk involved with Rubin's degrees of freedom is the possibility of obtaining degrees of freedom which are larger than the sample size. Adjustments to the degrees of freedom are presented by Barnard & Rubin (1999), Lipsitz et al. (2002), and Reiter (2007). These alternatives are compared in Wagstaff & Harel (2011).

Rates of Missing Information

It may be of interest to the researcher to examine the rates of missing information in multiple imputation. In general, to find the missing information using standard multiple imputation, one finds the posterior distribution of Q from the incomplete data and also the hypothetical Fisher information if there had been no missing values. The ratio of those two estimates is the rate of missing information. Harel (2007) derives an estimate for the rate of missing information as well as the asymptotic distribution. Let λ be the true rate of missing information. An estimate of λ is

$$\hat{\lambda} = \frac{B}{\bar{U} + B}$$

which does not change as the number of imputations increases. A full derivation of this estimate can be found in Harel (2007). Moreover, λ has an asymptotically normal distribution:

$$\frac{\sqrt{m}(\hat{\lambda} - \lambda)}{\sqrt{2}\lambda(1 - \lambda)} \rightarrow N(0, 1)$$

which allows us to calculate confidence intervals for λ.

1.2.2 Two-Stage Multiple Imputation

Conceptual Overview

Two-stage multiple imputation (or nested multiple imputation) involves generating imputations through a two step process to account for two different types of missing values. That is, consider a situation where the missing data are of two different types. These two types could be dropout in a longitudinal study and intermittent missed measurements. Another example is planned and unplanned values in survey studies. A third example could be missing responses and missing covariates. The possibilities are endless! If the missing data are of two different types, it may be beneficial to treat them differently and adjust the multiple imputation method accordingly.

The general idea behind two-stage multiple imputation is to impute the first type of missing values m times. Then, for each of those m data sets, impute the second type of missing values n times, treating the imputations for the first type of missing values as fixed and known. This yields a total of mn complete data sets from which estimates and variances are obtained (Harel, 2009).

Two types of missing values do not necessarily need to be treated separately. However, there are some advantages to this two-stage structure. One advantage is the ability to account for the degree of uncertainty contributed by each type of missing value. The second advantage is the ability to make different probabilistic assumptions on each of the types of missing values. This second advantage is particularly important because it allows the imputer to differentiate between missing values strongly related to the outcome being measured and values that are missing for some other reason irrelevant to the outcome.

Suppose the missing data are partitioned into two parts so that $Y = (Y_{obs}, Y^A_{mis}, Y^B_{mis})$. Define M^+ as a random variable which is a matrix of missing data indicators of the same size as Y_{com} with 0 in each position corresponding to Y_{obs}, 1 in each position corresponding to Y^A_{mis} and 2 in each position corresponding to Y^B_{mis}. Then, to carry out the two-stage multiple imputation, first draw m values of Y^A_{mis} from

$$Y^{A(j)}_{mis} \sim P(Y^A_{mis} | Y_{obs}, M^+) \tag{1.5}$$

and then for each $Y_{mis}^{A(j)}$ draw n values of Y_{mis}^{B} from

$$Y_{mis}^{B(j,k)} \sim P(Y_{mis}^{B}|Y_{obs}, Y_{mis}^{A(j)}, M^{+}). \tag{1.6}$$

Under certain ignorability conditions (Harel, 2009; Harel & Schafer, 2009), Equations (1.5) and (1.6) can be reduced to $P(Y_{mis}^{A}|Y_{obs})$ and $P(Y_{mis}^{B}|Y_{obs}, Y_{mis}^{A})$.

Large Sample Case

In an unpublished dissertation, Shen (2000) derived combining rules for this nested imputation scheme where he used the context of one set of imputations being computationally expensive and the other as being computationally inexpensive. Regardless of the application, the combining rules are reminiscent of a classical nested analysis of variance (ANOVA) with Y_{mis}^{A} as a blocking factor (Shen, 2000).

The overall point estimate is

$$\bar{Q} = \frac{1}{mn}\sum_{j=1}^{m}\sum_{k=1}^{n}\hat{Q}^{(j,k)} = \frac{1}{m}\sum \bar{Q}_{j.}$$

where $\bar{Q}_{j.}$ is the average of the m^{th} nest, which is similar in concept to the combining rule from standard multiple imputation. The variance of $\bar{Q}$, however, now has three components. There is the estimated complete data variance

$$\bar{U} = \frac{1}{mn} \sum_{j=1}^{m} \sum_{k=1}^{n} U^{(j,k)},$$

the between-block imputation variance

$$B = \frac{1}{m-1} \sum_{j=1}^{m} (\bar{Q}_{j.} - \bar{Q})^2,$$

and the within-block imputation variance

$$W = \frac{1}{m} \sum_{j=1}^{m} \frac{1}{n-1} \sum_{k=1}^{n} (\hat{Q}^{(j,k)} - \bar{Q}_{j.})^2.$$

Then, Shen (2000) was able to derive that the total variance is

$$T = \bar{U} + (1 + m^{-1})B + (1 - n^{-1})W$$

and from there could determine that inferences for Q should be based on a t-distribution; $(Q - \bar{Q})/\sqrt{T} \sim t_{\nu_*}$ with degrees of freedom

$$\nu_*^{-1} = \frac{1}{m-1} \left(\frac{(1 + 1/m)B}{T} \right)^2 + \frac{1}{m(n-1)} \left(\frac{(1 - 1/n)W}{T} \right)^2.$$

In a manner similar to equation (1.1), the variance of $Q|Y_{obs}$ can be expressed as follows:

$$
\begin{aligned}
V(Q|Y_{obs}) = {} & V(E(E(Q|Y_{obs}, Y_{mis}^{A}, Y_{mis}^{B})|Y_{obs}, Y_{mis}^{A})|Y_{obs}) \\
+ {} & E(V(E(Q|Y_{obs}, Y_{mis}^{A}, Y_{mis}^{B})|Y_{obs}, Y_{mis}^{A})|Y_{obs}) \\
+ {} & E(E(V(Q|Y_{obs}, Y_{mis}^{A}, Y_{mis}^{B})|Y_{obs}, Y_{mis}^{A})|Y_{obs}) \\
= {} & B_{\infty} + W_{\infty} + \bar{U}_{\infty}.
\end{aligned}
$$

Ignorability conditions in two-stage multiple imputation are explored in Harel (2009).

Rates of Missing Information

When the missing data are of two types, the rate of missing information due to each type of missing value can be quantified. Harel (2007) derived estimates for the overall rate of missing information (λ) and the rates attributable to each type of missing value

(λ^A and $\lambda^{B|A}$) as follows:

$$\begin{aligned}
\hat{\lambda} &= \frac{B + (1 - n^{-1})W}{\bar{U} + B + (1 - n^{-1})W} \\
\hat{\lambda}^{B|A} &= \frac{W}{\bar{U} + W} \\
\hat{\lambda}^{A} &= \hat{\lambda} - \hat{\lambda}^{B|A}.
\end{aligned}$$

Harel (2007) also provides a derivation of the asymptotic distribution of the rates of missing information and shows them to be asymptotically normal.

Chapter 2

Three-Stage Multiple Imputation

2.1 Procedure for Three-Stage Multiple Imputation

The approach used for implementing multiple imputation in three stages begins with partitioning the data. Let Y_{com} represent the complete data set. This set can be partitioned into four parts: $(Y_{obs}, Y_{mis}^A, Y_{mis}^B, Y_{mis}^C)$ where Y_{obs} represents the observed part of the data, Y_{mis}^A represents the part of the data missing due to the first type of missingness, Y_{mis}^B represents the part of the data missing due to the second type of missingness and Y_{mis}^C represents the part of the data missing due to the third type of missingness. Let M^+ represent the set of extended missingness indicators. M^+ is an array the same size as Y_{com} which contains a 0 in every position corresponding to an observed value, a 1 in every position corresponding to a value in Y_{mis}^A, a 2 in every position corresponding to a value in Y_{mis}^B, and a 3 in every position corresponding to a value in Y_{mis}^C. The joint model is now expressed in terms of M^+ and Y_{com}:

$$P(Y_{com}, M^+, \theta, \phi^+) = P(Y_{com}|\theta)P(M^+|Y_{com}, \phi^+)P(\theta, \phi^+) \tag{2.1}$$

where θ represents the parameter of interest and ϕ^+ represents the parameters of the extended missingness mechanism.

To implement three-stage multiple imputation, we use an extension of the procedure used in two-stage multiple imputation (Shen, 2000). First we draw L independent values of Y_{mis}^A from its predictive distribution,

$$Y_{mis}^{A(l)} \sim P(Y_{mis}^A | Y_{obs}, M^+), \quad l = 1, 2, \ldots, L. \tag{2.2}$$

Then, for each imputation of $Y_{mis}^{A(l)}$, we draw M conditionally independent values of Y_{mis}^B:

$$Y_{mis}^{B(l,m)} \sim P(Y_{mis}^B | Y_{obs}, Y_{mis}^{A(l)}, M^+), \quad m = 1, 2, \ldots, M. \tag{2.3}$$

Finally, for each imputation of $(Y_{mis}^{A(l)}, Y_{mis}^{B(l,m)})$ we draw N conditionally independent values of Y_{mis}^C:

$$Y_{mis}^{C(l,m,n)} \sim P(Y_{mis}^C | Y_{obs}, Y_{mis}^{A(l)}, Y_{mis}^{B(l,m)}, M^+), \quad n = 1, 2, \ldots, N \tag{2.4}$$

for a grand total of $L \times M \times N$ completed data sets.

Essentially, what is being done is drawing L conventional imputations of Y_{mis} then, for each of those, generating $(M-1)$ additional draws of Y_{mis}^B, treating Y_{mis}^A as fixed and finally, for each of those, generating $(N-2)$ additional draws of Y_{mis}^C, treating Y_{mis}^A and Y_{mis}^B as fixed. Under certain conditions, the information in M^+ can be ignored but those conditions will be discussed further in Chapter 4.

2.2 Bayesian Derivation

From a Bayesian point of view, the objective is to obtain the distribution of Q given the observed data, Y_{obs}, and M^+. Let $Y_{com}^{(l,m,n)} = \{Y_{obs}, Y_{mis}^{A(l)}, Y_{mis}^{B(l,m)}, Y_{mis}^{C(l,m,n)}\}$ be the complete data set. The posterior distribution of Q can be expressed as

$$\begin{aligned} P(Q|Y_{obs}, M^+) &= \int P(Q|Y_{obs}, Y_{mis}, M^+)P(Y_{mis}|Y_{obs}, M^+)dY_{mis} \\ &= \lim_{L\to\infty} \frac{1}{L}\sum_{l=1}^{L} P(Q|Y_{obs}, Y_{mis}^{A(l)}, Y_{mis}^{B(l,1)}, Y_{mis}^{C(l,1,1)}, M^+) \\ &= \lim_{L\to\infty} \frac{1}{L}\sum_{l=1}^{L} P(Q|Y_{com}^{(l,1,1)}). \end{aligned}$$

As with standard multiple imputation and two-stage multiple imputation, the goal is to use the completed data statistics to make inferences. Let

$$S_{LMN} = \{(\hat{Q}^{(l,m,n)}, U^{(l,m,n)}) : l = 1, ..., L, m = 1, ..., M, n = 1, ..., N\}$$

denote the set of completed data statistics.

Since the number of imputations, L, is finite, there is a loss of information using

S_{LMN} rather than Y_{obs}. The conditional distribution of Q given S_{LMN} and M^+ is

$$\begin{aligned} P(Q|S_{LMN}, M^+) &= \int P(Q|S_\infty, S_{LMN}, M^+)P(S_\infty|S_{LMN}, M^+)dS_\infty \\ &= \int P(Q|S_\infty, M^+)P(S_\infty|S_{LMN}, M^+)dS_\infty \end{aligned}$$

where S_∞ is S_{LMN} when $L = \infty$.

This conditional distribution is difficult to derive analytically so it is approximated in five steps.

1. Derive $P(Q|S_\infty, M^+)$. (Section 2.2.1)

2. Derive $P(\bar{Q}_\infty, \bar{U}_\infty|S_{LMN}, B_\infty^{(b)}, B_\infty^{(w1)}, B_\infty^{(w2)}, M^+)$ where these variables are functions of S_∞ to be defined in the derivation. (Section 2.2.2)

3. Combine results of the first two steps to get $P(Q|S_{LMN}, B_\infty^{(b)}, B_\infty^{(w1)}, B_\infty^{(w2)}, M^+)$. (Section 2.2.3)

4. Approximate $P(\bar{U}_{LMN} + \left(1 + \frac{1}{L}\right) B_\infty^{(b)} + \left(1 + \frac{1}{LM}\right) B_\infty^{(w1)} + \left(1 + \frac{1}{LMN}\right) B_\infty^{(w2)}|S_{LMN},$ $M^+)$. (Section 2.2.4)

5. Combine results of steps 3 and 4 to obtain an approximation to $P(Q|S_{LMN}, M^+)$. (Section 2.2.5)

The end goal is to approximate the posterior distribution of Q given the observed data.

2.2.1 Step 1

The first step of the derivation is to find the conditional distribution of Q given S_∞, M^+. Conditional on the completed data (Y_{obs}, Y_{mis}), the completed data posterior distribution of Q has mean $\hat{Q} = \hat{Q}(Y_{obs}, Y_{mis})$ and variance $U = U(Y_{obs}, Y_{mis})$. The actual posterior mean of Q is

$$E(Q|Y_{obs}, M^+) = E(\hat{Q}|Y_{obs}, M^+) \tag{2.5}$$

and the posterior variance of Q is

$$Var(Q|Y_{obs}, M^+) = E(U|Y_{obs}, M^+) + Var(\hat{Q}|Y_{obs}, M^+) \tag{2.6}$$

by the rules of iterative conditional expectation.

Suppose we have $L \times M \times N$ imputations where L is close to infinity. Then, we have $L \times M \times N$ sets of completed data statistics S_{LMN} as defined previously. The posterior mean of $\hat{Q}$ and the posterior mean of U are simulated in a manner similar to standard multiple imputation.

$$\bar{Q}_\infty \equiv \lim_{L\to\infty} \frac{1}{LMN} \sum_{l=1}^{L} \sum_{m=1}^{M} \sum_{n=1}^{N} \hat{Q}^{(l,m,n)} = E(\hat{Q}|Y_{obs}, M^+) \tag{2.7}$$

$$\bar{U}_\infty \equiv \lim_{L\to\infty} \frac{1}{LMN} \sum_{l=1}^{L} \sum_{m=1}^{M} \sum_{n=1}^{N} \hat{U}^{(l,m,n)} = E(\hat{U}|Y_{obs}, M^+). \tag{2.8}$$

Since the $L \times M \times N$ imputations are not independent, we must be cautious about simulating $Var(\hat{Q}|Y_{obs}, M^{+})$. Using Barnard (1995) orthogonal decomposition, $\hat{Q}$ can be decomposed into four orthogonal pieces.

$$\begin{aligned}
\hat{Q}(Y_{obs}, Y_{mis}^{A(l)}, Y_{mis}^{B(l,m)}, Y_{mis}^{C(l,m,n)}, M^{+}) &= H_0(Y_{obs}, M^{+}) \\
&+ H_1(Y_{mis}^{A(l)}, Y_{obs}, M^{+}) \\
&+ H_2(Y_{mis}^{A(l)}, Y_{mis}^{B(l,m)}, Y_{obs}, M^{+}) \\
&+ H_3(Y_{mis}^{A(l)}, Y_{mis}^{B(l,m)}, Y_{mis}^{C(l,m,n)}, Y_{obs}, M^{+})
\end{aligned} \tag{2.9}$$

where

$$\begin{aligned}
H_0(Y_{obs}, M^{+}) &= E(\hat{Q}|Y_{obs}, M^{+}) \\
H_1(Y_{obs}, Y_{mis}^{A(l)}, M^{+}) &= E(\hat{Q}|Y_{obs}, Y_{mis}^{A(l)}, M^{+}) - H_0 \\
H_2(Y_{obs}, Y_{mis}^{A(l)}, Y_{mis}^{B(l,m)}, M^{+}) &= E(\hat{Q}|Y_{obs}, Y_{mis}^{A(l)}, Y_{mis}^{B(l,m)}, M^{+}) - H_1 \\
H_3(Y_{obs}, Y_{mis}^{A(l)}, Y_{mis}^{B(l,m)}, Y_{mis}^{C(l,m,n)}, M^{+}) &= E(\hat{Q}|Y_{obs}, Y_{mis}^{A(l)}, Y_{mis}^{B(l,m)}, Y_{mis}^{C(l,m,n)} M^{+}) - H_2.
\end{aligned}$$

Using this construction, all of the pairwise correlations are zero. $Var(\hat{Q}|Y_{obs}, M^{+})$ can be expressed as

$$Var(\hat{Q}|Y_{obs}, M^{+}) = Var(H_1|Y_{obs}, M^{+}) + Var(H_2|Y_{obs}, M^{+}) + Var(H_3|Y_{obs}, M^{+}). \tag{2.10}$$

Table 1: Analysis of variance table

Source	Degree of Freedom	Sum of Squares	Mean Squares	Expected Mean Squares
Between nest	$L-1$	$MN\sum\left(\bar{Q}_{l..}-\bar{Q}_{...}\right)^2$	MS_L (MNB)	$\sigma^2_{W_2}+N\sigma^2_{W_1}+MN\sigma^2_B$
Within nest	$L(M-1)$	$N\sum\sum\left(\bar{Q}_{lm.}-\bar{Q}_{l..}\right)^2$	$MS_{M(L)}$ (NW_1)	$\sigma^2_{W_2}+N\sigma^2_{W_1}$
Within sub-nest	$LM(N-1)$	$\sum\sum\sum\left(\hat{Q}^{(l,m,n)}-\bar{Q}_{lm.}\right)^2$	$MS_{N(M)}$ (W_2)	$\sigma^2_{W_2}$

Using the ANOVA results (Table 1) for three-stage nested design,

$$
\begin{aligned}
Var(H_3|Y_{obs},M^+) &= B_\infty^{(w2)} \equiv \lim_{L\to\infty} MS_{N(M)}\\
Var(H_2|Y_{obs},M^+) &= B_\infty^{(w1)} \equiv \lim_{L\to\infty} \frac{MS_{M(L)}-MS_{N(M)}}{N}\\
Var(H_1|Y_{obs},M^+) &= B_\infty^{(b)} \quad \equiv \lim_{L\to\infty} \frac{MS_L-MS_{M(L)}}{MN}.
\end{aligned}
$$

Therefore, the posterior variance of $\hat{Q}$ is

$$Var(\hat{Q}|Y_{obs},M^+) = B_\infty^{(b)}+B_\infty^{(w1)}+B_\infty^{(w2)} = B_\infty. \tag{2.11}$$

From equations (2.5) and (2.7), we conclude that

$$E(Q|Y_{obs},M^+) = \bar{Q}_\infty \tag{2.12}$$

and from equations (2.6), (2.7), and (2.11),

$$Var(Q|Y_{obs}, M^{+}) = \bar{U}_{\infty} + B_{\infty} = T_{\infty}. \tag{2.13}$$

Then, assuming an approximate normal posterior distribution, we have

$$(Q|Y_{obs}, M^{+}) \sim N(\bar{Q}_{\infty}, T_{\infty}). \tag{2.14}$$

Under the normal posterior distribution, this posterior distribution is equivalent to the conditional distribution of Q given $h(S_{\infty}) = (\bar{Q}_{\infty}, \bar{U}_{\infty}, B_{\infty}^{(b)}, B_{\infty}^{(w1)}, B_{\infty}^{(w2)})$ which is what we wanted to show.

2.2.2 Step 2

The second step of the derivation is to find the distribution of

$$(\bar{Q}_{\infty}, \bar{U}_{\infty})|(S_{LMN}, B_{\infty}^{(b)}, B_{\infty}^{(w1)}, B_{\infty}^{(w2)}, M^{+}). \tag{2.15}$$

In order to do this, we must first find the conditional distribution of S_{LMN} given $(\bar{Q}_{\infty}, \bar{U}_{\infty}, B_{\infty}^{(b)}, B_{\infty}^{(w1)}, B_{\infty}^{(w2)})$ and then treat $(\bar{Q}_{\infty}, \bar{U}_{\infty})$ as estimands and S_{LMN} as data. Then, by assuming a flat prior on $\bar{Q}_{\infty}$, we can obtain the posterior distribution of $(\bar{Q}_{\infty}, \bar{U}_{\infty})$.

From the orthogonal decomposition (2.9), it can be shown that the covariance between any two $\hat{Q}$'s is as follows: $Cov(\hat{Q}^{(l_1,m_1,n_1)}, \hat{Q}^{(l_2,m_2,n_2)}|Y_{obs}, M^+)$

$$= \begin{cases} 0 : & l_1 \neq l_2 \\ B_\infty^{(b)} : & l_1 = l_2, m_1 \neq m_2 \\ B_\infty^{(b)} + B_\infty^{(w1)} : & l_1 = l_2, m_1 = m_2, n_1 \neq n_2 \\ B_\infty^{(b)} + B_\infty^{(w1)} + B_\infty^{(w2)} : & l_1 = l_2, m_1 = m_2, n_1 = n_2. \end{cases}$$

Let $\hat{Q}^{(l)}$ denote all the $\hat{Q}$'s in the l^{th} nest. That is,

$$\hat{Q}^{(l)} = (\hat{Q}^{(l,1,1)}, \hat{Q}^{(l,1,2)}, \ldots, \hat{Q}^{(l,M,N)})^T.$$

The variance-covariance matrix of $\hat{Q}^{(l)}$ conditioned on $\bar{Q}_\infty, \bar{U}_\infty, B_\infty^{(b)}, B_\infty^{(w1)}$, and $B_\infty^{(w2)}$ can be expressed as

$$Var(\hat{Q}^{(l)}|Y_{obs}, M^+) = B_\infty^{(b)} 1_{MN\times MN} + \bigoplus_{i=1}^{M} [B_\infty^{(w1)} 1_{N\times N} + B_\infty^{(w2)} I_{N\times N}] \qquad (2.16)$$

where $1_{MN\times MN}$ and $1_{N\times N}$ are square matrices with all the elements 1 and $I_{N\times N}$ is the identity matrix with dimension $N \times N$. The $\bigoplus_{i=1}^{M}$ represents the direct sum of the matrices which creates a block diagonal matrix.

The conditional expectation of each element in $\hat{Q}^{(l)}$ is

$$E(\hat{Q}^{(l,m,n)}|Y_{obs}, M^+) = \bar{Q}_\infty. \tag{2.17}$$

Note that vectors from different nests are independent. That is, $\hat{Q}^{(l_1)}$ is independent of $\hat{Q}^{(l_2)}$. With the assumption of large-sample data, we can assume that given (Y_{obs}, M^+), $\hat{Q}^{(l)}$ are i.i.d. draws from

$$(\hat{Q}^{(l)}|\bar{Q}_\infty, B_\infty^{(b)}, B_\infty^{(w1)}, B_\infty^{(w2)}) \sim N(\bar{Q}_\infty 1_{MN}, B_\infty^{(b)} 1_{MN\times MN} + \bigoplus_{i=1}^{M} [B_\infty^{(w1)} 1_{N\times N} + B_\infty^{(w2)} I_{N\times N}]). \tag{2.18}$$

Then the conditional expectation and variance follow:

$$\begin{aligned} E(\bar{Q}_{LMN}|Y_{obs}, M^+) &= E\left(\frac{1}{LMN}\sum_{l=1}^{L}\sum_{m=1}^{M}\sum_{n=1}^{N}\hat{Q}^{(l,m,n)}\right) \\ &= \bar{Q}_\infty \end{aligned} \tag{2.19}$$

and

$$\begin{aligned} Var(\bar{Q}_{LMN}|Y_{obs}, M^+) &= Var\left(\frac{1}{LMN}\sum_{l=1}^{L}\sum_{m=1}^{M}\sum_{n=1}^{N}\hat{Q}^{(l,m,n)}\right) \\ &= \frac{1}{L}B_\infty^{(b)} + \frac{1}{LM}B_\infty^{(w1)} + \frac{1}{LMN}B_\infty^{(w2)}. \end{aligned} \tag{2.20}$$

Since $\bar{Q}_{LMN}$ is a linear combination of jointly normal random variables, its sampling distribution also follows a normal distribution of the form

$$(\bar{Q}_{LMN}|\bar{Q}_{\infty}, B_{\infty}^{(b)}, B_{\infty}^{(w1)}, B_{\infty}^{(w2)}, M^{+}) \sim N\left(\bar{Q}_{\infty}, \frac{1}{L}B_{\infty}^{(b)} + \frac{1}{LM}B_{\infty}^{(w1)} + \frac{1}{LMN}B_{\infty}^{(w2)}\right). \tag{2.21}$$

The joint distribution of $U^{(l,m,n)}$ is much harder to specify but it is not necessary to do so. Instead, we use the assumption of lower order of variability in the same way as standard multiple imputation:

$$(U^{(l,m,n)}|Y_{obs}, M^{+}) \sim (\bar{U}_{\infty}, \ll B_{\infty}) \tag{2.22}$$

where $D \sim (A, \ll C)$ means that the distribution of D is centered around A with each component having variability substantially less than each positive component of C.

Now, we treat $(\bar{Q}_{\infty}, \bar{U}_{\infty})$ as two unknown estimands with a conditional distribution that can be found by

$$\begin{aligned} P(\bar{Q}_{\infty}, \bar{U}_{\infty}|S_{LMN}, B_{\infty}^{(b)}, B_{\infty}^{(w1)}, B_{\infty}^{(w2)}, M^{+}) \propto \\ P(\bar{Q}_{\infty}, \bar{U}_{\infty})P(S_{LMN}|\bar{Q}_{\infty}, \bar{U}_{\infty}, B_{\infty}^{(b)}, B_{\infty}^{(w1)}, B_{\infty}^{(w2)}, M^{+}). \end{aligned} \tag{2.23}$$

Assume a flat prior for $\bar{Q}_{\infty}$ and combine with the sampling distribution of $\bar{Q}_{LMN}$ to get

$$(\bar{Q}_\infty | S_{LMN}, B_\infty^{(b)}, B_\infty^{(w1)}, B_\infty^{(w2)}, M^+) \sim N\left(\bar{Q}_{LMN}, \frac{1}{L}B_\infty^{(b)} + \frac{1}{LM}B_\infty^{(w1)} + \frac{1}{LMN}B_\infty^{(w2)}\right). \tag{2.24}$$

Also, assuming a relatively diffuse prior on $\bar{U}_\infty$ combined with the sampling distribution of $U^{(l,m,n)}$, we get

$$(\bar{U}_\infty | S_{LMN}, B_\infty^{(b)}, B_\infty^{(w1)}, B_\infty^{(w2)}, M^+) \sim \left(\bar{U}_{LMN}, \ll \frac{B_\infty}{LMN}\right). \tag{2.25}$$

2.2.3 Step 3

In this step of the derivation, the results from the previous two steps are combined to get the conditional distribution of Q given $(S_{LMN}, B_\infty^{(b)}, B_\infty^{(w1)}, B_\infty^{(w2)}, M^+)$. This is achieved with the following integration:

$$\begin{aligned}
&P(Q|S_{LMN}, B_\infty^{(b)}, B_\infty^{(w1)}, B_\infty^{(w2)}, M^+) && (2.26)\\
&\quad = \int P(Q|\bar{Q}_\infty, \bar{U}_\infty, B_\infty^{(b)}, B_\infty^{(w1)}, B_\infty^{(w2)}, S_{LMN}, M^+) \\
&\qquad \times P(\bar{Q}_\infty, \bar{U}_\infty | B_\infty^{(b)}, B_\infty^{(w1)}, B_\infty^{(w2)}, S_{LMN}, M^+) d\bar{Q}_\infty d\bar{U}_\infty \\
&\quad = \int P(Q|\bar{Q}_\infty, \bar{U}_\infty, B_\infty^{(b)}, B_\infty^{(w1)}, B_\infty^{(w2)}, M^+) \\
&\qquad \times P(\bar{Q}_\infty, \bar{U}_\infty | B_\infty^{(b)}, B_\infty^{(w1)}, B_\infty^{(w2)}, S_{LMN}, M^+) d\bar{Q}_\infty d\bar{U}_\infty. && (2.27)
\end{aligned}$$

Note that the distribution of $(Q|\bar{Q}_\infty, \bar{U}_\infty, B_\infty^{(b)}, B_\infty^{(w1)}, B_\infty^{(w2)}, S_{LMN}, M^+)$ is $\mathrm{N}(\bar{Q}_\infty, \bar{U}_\infty + B_\infty)$ which is replaced with $\mathrm{N}(\bar{Q}_\infty, \bar{U}_{LMN} + B_\infty)$ because of the lower variability result. Combined with the conditional distribution of $\bar{Q}_\infty$, we get

$$(Q|S_{LMN}, B_\infty^{(b)}, B_\infty^{(w1)}, B_\infty^{(w2)}, M^+) \sim N\left(\bar{Q}_{LMN}, \bar{U}_{LMN} + \left(1+\frac{1}{L}\right)B_\infty^{(b)} + \left(1+\frac{1}{LM}\right)B_\infty^{(w1)} + \left(1+\frac{1}{LMN}\right)B_\infty^{(w2)}\right). \quad (2.28)$$

2.2.4 Step 4

The plan for this step of the derivation is to approximate the variance, which is a function of $B_\infty^{(b)}, B_\infty^{(w1)}$, and $B_\infty^{(w2)}$, by an inverse χ^2 distribution.

Define the expected mean squares as

$$\begin{aligned} ES^{(b)} &= E(MS_L) \\ ES^{(w1)} &= E(MS_{M(L)}) \\ ES^{(w2)} &= E(MS_{N(M)}). \end{aligned} \quad (2.29)$$

Assuming the improper prior

$$\pi(ES^{(b)}, ES^{(w1)}, ES^{(w2)}) \propto \frac{1}{ES^{(b)}} \times \frac{1}{ES^{(w1)}} \times \frac{1}{ES^{(w2)}} \quad (2.30)$$

and the ANOVA table, we have

$$
\begin{aligned}
ES^{(b)} &\sim \chi^{-2}(L-1, MS_L) \\
ES^{(w1)} &\sim \chi^{-2}(L(M-1), MS_{M(L)}) \\
ES^{(w2)} &\sim \chi^{-2}(LM(N-1), MS_{N(M)})
\end{aligned}
\tag{2.31}
$$

where $\chi^{-2}(\nu, s)$ denotes a scaled inverse χ^2 distribution with degree of freedom ν and scale s.

Also from the ANOVA table, we have

$$
\begin{aligned}
ES^{(b)} &= B_{\infty}^{(w2)} + NB_{\infty}^{(w1)} + MNB_{\infty}^{(b)} \\
ES^{(w1)} &= B_{\infty}^{(w2)} + NB_{\infty}^{(w1)} \\
ES^{(w2)} &= B_{\infty}^{(w2)}.
\end{aligned}
\tag{2.32}
$$

Now the conditional variance of Q given $(S_{LMN}, B_{\infty}^{(b)}, B_{\infty}^{(w1)}, B_{\infty}^{(w2)}, M^{+})$ can be rewritten as

$$
\frac{1}{MN}\left(1+\frac{1}{L}\right) ES^{(b)} + \frac{1}{N}\left(1-\frac{1}{M}\right) ES^{(w1)} + \left(1-\frac{1}{N}\right) ES^{(w2)}. \tag{2.33}
$$

Note that if $N = 1$, the above equation reduces to the form for two-stage multiple imputation and if $N = M = 1$, the above equation reduces to the form for standard multiple imputation.

The goal is to rewrite the conditional variance of Q as a linear combination of three scaled inverse χ^2's plus a constant and then integrate out $ES^{(b)}, ES^{(w1)}$, and $ES^{(w2)}$ over the conditional distribution of $(ES^{(b)}, ES^{(w1)}, ES^{(w2)}|S_{LMN}, M^+)$. However, this is challenging given the constraint $ES^{(b)} \geq ES^{(w1)} \geq ES^{(w2)}$. Instead, the approach from standard multiple imputation is used and we approximate the conditional variance by a scaled inverse χ^2. By relaxing the constraint and assuming that $ES^{(b)}, ES^{(w1)}$, and $ES^{(w2)}$ are independent scaled inverse χ^2's, we can match the first two moments. From the ANOVA table, we have

$$\begin{aligned} MS_L &\sim ES^{(b)}\chi^2[L-1]/(L-1) \\ MS_{M(L)} &\sim ES^{(w1)}\chi^2[L(M-1)]/(L(M-1)) \\ MS_{N(M)} &\sim ES^{(w2)}\chi^2[LM(N-1)]/(LM(N-1)). \end{aligned} \tag{2.34}$$

For convenience, a conjugate prior distribution is chosen for $ES^{(b)}, ES^{(w1)}, ES^{(w2)}$

and has the form:

$$
\begin{aligned}
\pi(ES^{(b)}, ES^{(w1)}, ES^{(w2)} | r^{(b)}, r^{(w1)}, r^{(w2)}, \omega^{(b)}, \omega^{(w1)}, \omega^{(w2)}) \\
\propto (ES^{(b)})^{-(\frac{\omega^{(b)}}{2}+1)} exp\left[-\frac{r^{(b)}\omega^{(b)}}{2ES^{(b)}}\right] \\
\times (ES^{(w1)})^{-(\frac{\omega^{(w1)}}{2}+1)} exp\left[-\frac{r^{(w1)}\omega^{(w1)}}{2ES^{(w1)}}\right] \\
\times (ES^{(w2)})^{-(\frac{\omega^{(w2)}}{2}+1)} exp\left[-\frac{r^{(w2)}\omega^{(w2)}}{2ES^{(w2)}}\right]
\end{aligned} \tag{2.35}
$$

where $r^{(b)}$, $\omega^{(b)}$, $r^{(w1)}$, $\omega^{(w1)}$, $r^{(w2)}$, and $\omega^{(w2)}$ are the parameters corresponding to inverse gamma distributions for $ES^{(b)}$, $ES^{(w1)}$, and $ES^{(w2)}$, respectively.

Then, combining equations (2.34) and (2.35), we get

$$
\begin{aligned}
&\pi(ES^{(b)}, ES^{(w1)}, ES^{(w2)} | S_{LMN}) \propto \\
&(ES^{(b)})^{-(\frac{(L-1)+\omega^{(b)}}{2}+1)} exp\left[-\frac{(L-1)MS_L + r^{(b)}\omega^{(b)}}{2ES^{(b)}}\right] \\
&\times (ES^{(w1)})^{-(\frac{L(M-1)+\omega^{(w1)}}{2}+1)} exp\left[-\frac{L(M-1)MS_{M(L)} + r^{(w1)}\omega^{(w1)}}{2ES^{(w1)}}\right] \\
&\times (ES^{(w2)})^{-(\frac{LM(N-1)+\omega^{(w2)}}{2}+1)} exp\left[-\frac{LM(N-1)MS_{N(M)} + r^{(w2)}\omega^{(w2)}}{2ES^{(w2)}}\right]
\end{aligned} \tag{2.36}
$$

where $ES^{(b)} \leq ES^{(w1)} \leq ES^{(w2)}$. Without that constraint, $ES^{(b)}, ES^{(w1)}$, and $ES^{(w2)}$ will be independently distributed as scaled inverse χ^2 distributions. $ES^{(b)}$ follows a scaled inverse χ^2 with degrees of freedom $\nu_1 = L-1+\omega^{(b)}$ and scale $s_1 = (L-1)MS_L + r^{(b)}\omega^{(b)}$ and $ES^{(w1)}$ follows a scaled inverse χ^2 with degrees of freedom $\nu_2 = L(M-1) + \omega^{(w1)}$

and scale $s_2 = L(M-1)MS_{M(L)} + r^{(w1)}\omega^{(w1)}$ and $ES^{(w2)}$ follows a scaled inverse χ^2 with degrees of freedom $\nu_3 = LM(N-1) + \omega^{(w2)}$ and scale $s_3 = LM(N-1)MS_{N(M)} + r^{(w2)}\omega^{(w2)}$.

Suppose $x_i^{-1}, i = 1, \ldots, p$ is independently distributed as a mean square random variable with f_i degrees of freedom. Then the quantity

$$D = \frac{1 + \sum_{i=1}^{p} a_i}{1 + \sum_{i=1}^{p} a_i x_i} \tag{2.37}$$

is approximately distributed as a mean square random variable with degrees of freedom

$$\nu^{-1} = \sum_{i=1}^{p} \left[\frac{a_i}{1 + \sum_{i=1}^{p} a_i} \right]^2 \frac{1}{f_i}. \tag{2.38}$$

Let

$$\begin{aligned}
x_1 &= \frac{ES^{(b)}(L - 1 + \omega^{(b)})}{(L-1)MS_L + r^{(b)}\omega^{(b)}}, \\
x_2 &= \frac{ES^{(w1)}(L(M-1) + \omega^{(w1)})}{L(M-1)MS_{M(L)} + r^{(w1)}\omega^{(w1)}}, \\
x_3 &= \frac{ES^{(w2)}(LM(N-1) + \omega^{(w2)})}{LM(N-1)MS_{N(M)} + r^{(w2)}\omega^{(w2)}}.
\end{aligned} \tag{2.39}$$

Under the assumption that $ES^{(b)}, ES^{(w1)}$, and $ES^{(w2)}$ are independent, then x_1, x_2, and x_3 are also independent mean square random variables.

Now, rewrite

$$\bar{U}_{LMN} + \frac{1}{MN}\left(1+\frac{1}{L}\right)ES^{(b)} + \frac{1}{N}\left(1-\frac{1}{M}\right)ES^{(w1)} + \left(1-\frac{1}{N}\right)ES^{(w2)} \tag{2.40}$$

as

$$\bar{U}_{LMN}\left(1+\frac{\beta_1}{\bar{U}_{LMN}}x_1+\frac{\beta_2}{\bar{U}_{LMN}}x_2+\frac{\beta_3}{\bar{U}_{LMN}}x_3\right) \tag{2.41}$$

where

$$\begin{aligned}
\beta_1 &= \frac{1}{MN}\left(1+\frac{1}{L}\right)\frac{(L-1)MS_L + r^{(b)}\omega^{(b)}}{(L-1+\omega^{(b)})} \\
\beta_2 &= \frac{1}{N}\left(1-\frac{1}{M}\right)\frac{L(M-1)MS_{M(L)} + r^{(w1)}\omega^{(w1)}}{(L(M-1)+\omega^{(w1)})} \\
\beta_3 &= \left(1-\frac{1}{N}\right)\frac{LM(N-1)MS_{N(M)} + r^{(w2)}\omega^{(w2)}}{(LM(N-1)+\omega^{(w2)})}.
\end{aligned} \tag{2.42}$$

From the approximation as shown previously, we know that (2.41) can be approximated by a scaled inverse χ^2 distribution with scale

$$T_{LMN} = \bar{U}_{LMN} + \beta_1 + \beta_2 + \beta_3 \tag{2.43}$$

and degrees of freedom

$$\nu^{-1} = \left[\frac{\beta_1}{T_{LMN}}\right]^2 (L-1+\omega^{(b)})^{-1} + \left[\frac{\beta_2}{T_{LMN}}\right]^2 (L(M-1)+\omega^{(w1)})^{-1} + \left[\frac{\beta_3}{T_{LMN}}\right]^2 (LM(N-1)+\omega^{(w2)})^{-1}.$$

Therefore, it is concluded that

$$\bar{U}_{LMN} + \left(1+\frac{1}{L}\right) B_\infty^{(b)} + \left(1+\frac{1}{LM}\right) B_\infty^{(w1)} + \left(1+\frac{1}{LMN}\right) B_\infty^{(w2)} | S_{LMN} \sim \chi^{-2}(\nu, T_{LMN}). \tag{2.44}$$

Combined with the improper prior (2.30) and without the constraint that $ES^{(b)} \geq ES^{(w1)} \geq ES^{(w2)}$, the inverse χ^2 distribution has scale parameter

$$T_{LMN} = \bar{U}_{LMN} + \frac{1}{MN}\left(1+\frac{1}{L}\right) MS_L + \frac{1}{N}\left(1-\frac{1}{M}\right) MS_{M(L)} + \left(1-\frac{1}{N}\right) MS_{N(M)} \tag{2.45}$$

and degrees of freedom

$$\nu^{-1} = \left[\frac{\frac{1}{MN}\left(1+\frac{1}{L}\right) MS_L}{T_{LMN}}\right]^2 (L-1)^{-1} + \left[\frac{\frac{1}{N}\left(1-\frac{1}{M}\right) MS_{M(L)}}{T_{LMN}}\right]^2 (L(M-1))^{-1} + \left[\frac{\left(1-\frac{1}{N}\right) MS_{N(M)}}{T_{LMN}}\right]^2 (LM(N-1))^{-1}. \tag{2.46}$$

Note that if $N=1$ this reduces to the two-stage case and if $N=M=1$ this reduces to standard multiple imputation.

2.2.5 Step 5

The final step is to arrive at the conditional distribution of Q given S_{LMN}, M^+ by integrating the conditional distribution of

$$Q|\bar{U}_{LMN} + \left(1+\frac{1}{L}\right)B_{\infty}^{(b)} + \left(1+\frac{1}{LM}\right)B_{\infty}^{(w1)} + \left(1+\frac{1}{LMN}B_{\infty}^{(w2)}\right), S_{LMN}, M^+$$

over the conditional distribution in

$$\bar{U}_{LMN} + \left(1+\frac{1}{L}\right)B_{\infty}^{(b)} + \left(1+\frac{1}{LM}\right)B_{\infty}^{(w1)} + \left(1+\frac{1}{LMN}B_{\infty}^{(w2)}\right)|S_{LMN}, M^+$$

to get

$$(Q|S_{LMN}, M^+) \sim t_{\nu}(\bar{Q}_{LMN}, T_{LMN}). \tag{2.47}$$

The $100(1-\alpha)\%$ interval estimate for Q is

$$\left(\bar{Q}_{LMN} - t_{\nu}(100(1-\alpha/2))\sqrt{T_{LMN}}, \bar{Q}_{LMN} + t_{\nu}(100(1-\alpha/2))\sqrt{T_{LMN}}\right). \tag{2.48}$$

2.2.6 Alternative Notation

For consistency with the notation of Rubin (1987) and Harel (2009), the results can be expressed as follows:

$$
\begin{aligned}
\bar{Q} &= \frac{1}{LMN} \sum_{l=1}^{L} \sum_{m=1}^{M} \sum_{n=1}^{N} \hat{Q}^{(l,m,n)} \\
\bar{U} &= \frac{1}{LMN} \sum_{l=1}^{L} \sum_{m=1}^{M} \sum_{n=1}^{N} U^{(l,m,n)} \\
B &= \frac{1}{L-1} \sum_{l=1}^{L} \left(\bar{Q}_{l..} - \bar{Q}_{...}\right)^2 \\
W_1 &= \frac{1}{L(M-1)} \sum_{l=1}^{L} \sum_{m=1}^{M} \left(\bar{Q}_{lm.} - \bar{Q}_{l..}\right)^2 \\
W_2 &= \frac{1}{LM(N-1)} \sum_{l=1}^{L} \sum_{m=1}^{M} \sum_{n=1}^{N} \left(\hat{Q}^{(l,m,n)} - \bar{Q}_{lm.}\right)^2 .
\end{aligned}
$$

T_{LMN} can be expressed as

$$T = \bar{U}_{LMN} + \left(1 + \frac{1}{L}\right) B + \left(1 - \frac{1}{M}\right) W_1 + \left(1 - \frac{1}{N}\right) W_2 \tag{2.49}$$

and the degrees of freedom can be expressed as

$$
\begin{aligned}
\nu^{-1} = &\left[\frac{\left(1 + \frac{1}{L}\right) B}{T}\right]^2 (L-1)^{-1} + \left[\frac{\left(1 - \frac{1}{M}\right) W_1}{T}\right]^2 (L(M-1))^{-1} \\
&+ \left[\frac{\left(1 - \frac{1}{N}\right) W_2}{T}\right]^2 (LM(N-1))^{-1}.
\end{aligned} \tag{2.50}
$$

2.3 Simulations

In order to evaluate the effectiveness of the method proposed, simulations were run under varying scenarios to assess viability. Among the questions of interest are "Does the number of imputations significantly affect bias and coverage?" and "Does the order of imputation significantly affect bias and coverage?" All of the simulations in this section assume ignorability.

2.3.1 Simulation Set-Up

Data for the simulations are generated through the following steps. X_1 is generated from a $N(50, 100)$ and X_2 is generated from $N(20, 100)$. 100 values of X_1 and X_2 are randomly drawn. Y is created by taking $Y = 2X_1 + 3X_2$ so that Y has a known mean of 160. Then, Y_{com} is the data which includes Y and X_1. X_2 may be thought of as random error.

A missingness structure is imposed as follows:

1. The first type of missing value is created with an MCAR structure to simulate a missing covariate. That is, a prespecified percentage of the values in X_1 are randomly deleted. Let $MCAR\%$ denote the percentage of missing values due to the first type of missingness.

2. The second type of missing value is created under an MAR structure. Let $MAR_1\%$ denote the percentage of missing values due to the first type of MAR. Values in Y

are removed if they are above the top $MAR_1\%$ percentile of X_1.

3. The third type of missing value is created under an MAR structure. Let $MAR_2\%$ denote the percentage of missing values due to the second type of MAR. Values in Y are removed if they are below the bottom $MAR_2\%$ percentile of X_1.

The missing values are imputed using the norm package in R (Schafer & Novo, 2013) with varying numbers of imputations at each stage denoted by the ordered triple (L, M, N) and such that the order is $MCAR\%$, $MAR_1\%$, $MAR_2\%$.

The parameter of interest, Q, is the mean of Y. Data are generated, imputed, and analyzed 1000 times for percent bias, MSE, and coverage. The nominal coverage is 95%.

2.3.2 Number of Imputations

In order to establish how many imputations are necessary for minimal bias and coverage near 95%, simulations were performed varying the number of imputations and assessing the differences for different percentages of missing values. Table 2 displays the breakdown of the overall percentages of missing values by each of the three types. The results are presented in Table 3 and Figure 1. The table displays the estimates of the percent bias, the mean square error, and the coverage probability varying the number of imputations at each stage. Figure 1 graphically displays the change in coverage for the varying numbers of imputations for each of the four combinations of missing values.

The results of the simulation indicate that increasing the number of imputations

Table 2: Breakdown of percentages of missing values by each type (ignorable case)

Total Percentage of Missing Values	$MCAR\%$	$MAR_1\%$	$MAR_2\%$
75	25	25	25
50	15	15	20
40	15	5	20
15	5	5	5

does not significantly improve coverage. Figure 1 shows no distinct patterns of improved coverage as either the number of imputations or the percentage of missing values changes. Table 3 shows that the percent bias is less than 1% in all cases and that the MSE increases as the percentage of missing values increases, which is what one would expect. Therefore, increasing the number of imputations does not significantly improve the results and a small number of imputations can be used to produce efficient results.

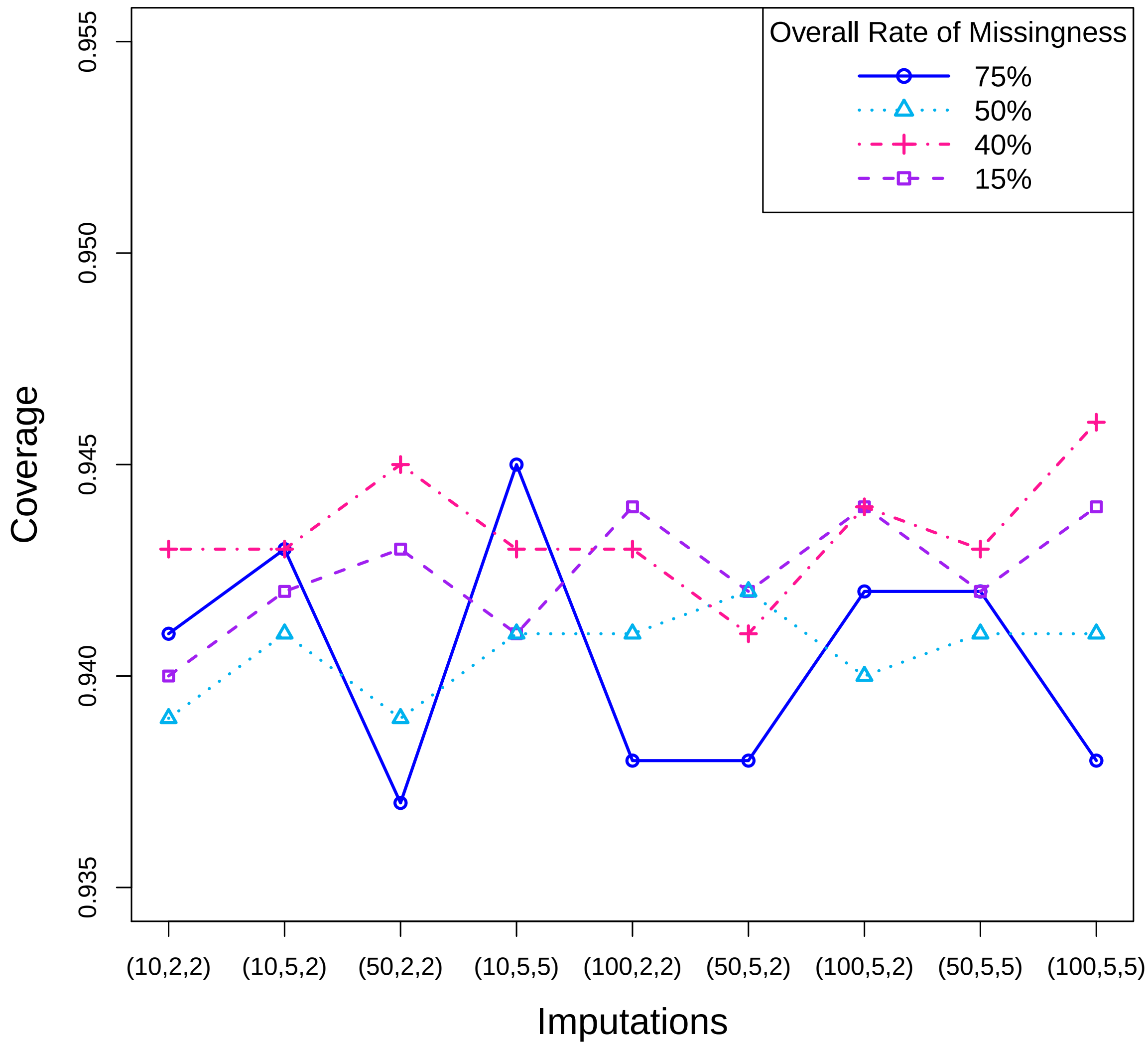

Figure 1: Estimates of coverage varying number of imputations at each stage for four combinations of percentages of missing values (ignorable case)

Percent Missing	(L,M,N)	Percent Bias (%)	MSE	Coverage
75	(10,2,2)	0.010	23.207	0.941
	(10,5,2)	0.024	23.296	0.943
	(10,5,5)	0.024	23.174	0.945
	(50,2,2)	0.007	23.178	0.937
	(50,5,2)	0.019	23.122	0.938
	(50,5,5)	0.018	23.125	0.942
	(100,2,2)	0.011	23.077	0.938
	(100,5,2)	0.021	23.042	0.942
	(100,5,5)	0.021	23.079	0.938
50	(10,2,2)	0.181	19.098	0.939
	(10,5,2)	0.141	18.912	0.941
	(10,5,5)	0.137	18.844	0.941
	(50,2,2)	0.173	18.919	0.939
	(50,5,2)	0.134	18.821	0.942
	(50,5,5)	0.131	18.812	0.941
	(100,2,2)	0.172	18.896	0.941
	(100,5,2)	0.134	18.817	0.940
	(100,5,5)	0.131	18.805	0.941
40	(10,2,2)	0.493	18.338	0.943
	(10,5,2)	0.487	18.141	0.943
	(10,5,5)	0.486	18.110	0.943
	(50,2,2)	0.498	18.071	0.945
	(50,5,2)	0.491	18.043	0.941
	(50,5,5)	0.488	18.027	0.943
	(100,2,2)	0.498	18.022	0.943
	(100,5,2)	0.492	18.002	0.944
	(100,5,5)	0.489	17.994	0.946
15	(10,2,2)	-0.112	15.081	0.940
	(10,5,2)	-0.113	15.052	0.942
	(10,5,5)	-0.112	15.074	0.941
	(50,2,2)	-0.113	15.108	0.943
	(50,5,2)	-0.111	15.093	0.942
	(50,5,5)	-0.110	15.094	0.942
	(100,2,2)	-0.112	15.080	0.944
	(100,5,2)	-0.110	15.086	0.944
	(100,5,5)	-0.109	15.084	0.944

Table 3: Estimates of percent bias, MSE, and coverage varying the percentages of missing values and number of imputations at each stage

2.3.3 Order of Imputation

In order to assess whether the order of imputation impacts the bias, coverage and MSE, a second simulation was run under the same structure as the first with the only difference being that the missing values were imputed in the following order: $MAR_2\%$, $MAR_1\%$, $MCAR\%$. That is, the missing values in the covariate were imputed last instead of first. Table 4 and Figure 2 display the results of the simulation. For simplicity, the table refers to the two different orders as $MCAR_F$ and $MCAR_L$ to mean the simulation where $MCAR\%$ is imputed first and the simulation where $MCAR\%$ is imputed last, respectively. The figure differentiates the two by using a solid line to represent the simulation where $MCAR\%$ is imputed first and a dotted line to represent the simulation where $MCAR\%$ is imputed last.

The results show that there is not a consistent change in the bias, MSE, or coverage when the order of imputation is switched. However, there appears to be a larger range of coverage values for the case where $MCAR\%$ is imputed last. The figure shows no discernible pattern that prefers one order over another and the table shows that neither the bias nor the MSE are consistently higher for one order over another. Therefore, in the situation where there is ignorability, the order of imputation does not appear to matter significantly.

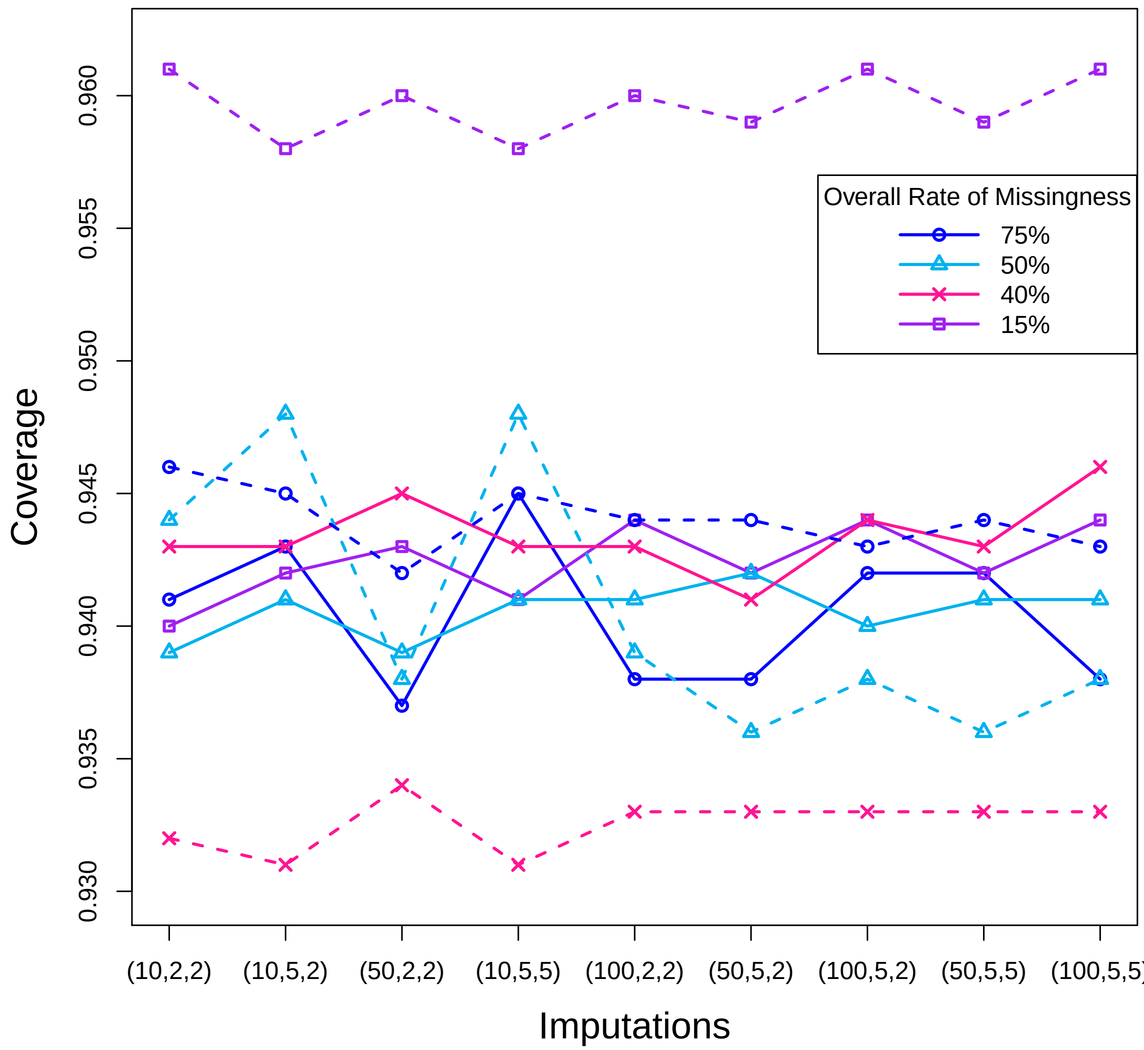

Figure 2: Estimates of coverage varying number of imputations at each stage for each of the four combinations of percentages of missing values. The solid lines are used to represent the simulations where $MCAR\%$ is imputed first and the dotted lines to represent the simulations where $MCAR\%$ is imputed last

Percent Missing	(L,M,N)	Percent Bias (%)		MSE		Coverage	
		$MCAR_F$	$MCAR_L$	$MCAR_F$	$MCAR_L$	$MCAR_F$	$MCAR_L$
75	(10,2,2)	0.010	0.009	23.207	24.815	0.941	0.946
	(10,5,2)	0.024	0.011	23.296	24.327	0.943	0.945
	(10,5,5)	0.024	0.011	23.174	24.327	0.945	0.945
	(50,2,2)	0.007	-0.002	23.178	24.210	0.937	0.942
	(50,5,2)	0.019	0.008	23.122	24.048	0.938	0.944
	(50,5,5)	0.018	0.008	23.125	24.048	0.942	0.944
	(100,2,2)	0.011	0.003	23.077	24.000	0.938	0.944
	(100,5,2)	0.021	0.011	23.042	23.930	0.942	0.943
	(100,5,5)	0.021	0.011	23.079	23.930	0.938	0.943
50	(10,2,2)	0.181	0.307	19.098	19.920	0.939	0.944
	(10,5,2)	0.141	0.304	18.912	19.827	0.941	0.948
	(10,5,5)	0.137	0.304	18.844	19.827	0.941	0.948
	(50,2,2)	0.173	0.285	18.919	19.930	0.939	0.938
	(50,5,2)	0.134	0.281	18.821	19.877	0.942	0.936
	(50,5,5)	0.131	0.281	18.812	19.877	0.941	0.936
	(100,2,2)	0.172	0.286	18.896	19.941	0.941	0.939
	(100,5,2)	0.134	0.281	18.817	19.906	0.940	0.938
	(100,5,5)	0.131	0.281	18.805	19.906	0.941	0.938
40	(10,2,2)	0.493	0.586	18.338	19.175	0.943	0.932
	(10,5,2)	0.487	0.581	18.141	19.144	0.943	0.931
	(10,5,5)	0.486	0.581	18.110	19.144	0.943	0.931
	(50,2,2)	0.498	0.580	18.071	18.792	0.945	0.934
	(50,5,2)	0.491	0.576	18.043	18.763	0.941	0.933
	(50,5,5)	0.488	0.576	18.027	18.763	0.943	0.933
	(100,2,2)	0.498	0.576	18.022	18.644	0.943	0.933
	(100,5,2)	0.492	0.573	18.002	18.623	0.944	0.933
	(100,5,5)	0.489	0.573	17.994	18.623	0.946	0.933
15	(10,2,2)	-0.112	0.051	15.081	13.189	0.940	0.961
	(10,5,2)	-0.113	0.058	15.052	13.156	0.942	0.958
	(10,5,5)	-0.112	0.058	15.074	13.156	0.941	0.958
	(50,2,2)	-0.113	0.052	15.108	13.116	0.943	0.960
	(50,5,2)	-0.111	0.056	15.093	13.118	0.942	0.959
	(50,5,5)	-0.110	0.056	15.094	13.118	0.942	0.959
	(100,2,2)	-0.112	0.051	15.080	13.124	0.944	0.960
	(100,5,2)	-0.110	0.055	15.086	13.121	0.944	0.961
	(100,5,5)	-0.109	0.055	15.084	13.121	0.944	0.961

Table 4: Comparison of estimates of percent bias, MSE, and coverage varying the percentages of missing values and number of imputations at each stage changing the order of imputation. For simplicity, the table refers to the two different orders as $MCAR_F$ and $MCAR_L$ to mean the simulation where $MCAR\%$ is imputed first and the simulation where $MCAR\%$ is imputed last, respectively

Chapter 3

Rates of Missing Information

This chapter describes the estimates and distributions of rates of missing information for three-stage multiple imputation. The rates of missing information can be important in quantifying how much missing information is contributing to uncertainty of Q and in deciding how many imputations should be used at each stage.

3.1 Estimation of Missing Information Rates

If the missing data carried no information about Q, then the imputed data estimates would be identical and T would reduce to $\bar{U}$. The information in the posterior predictive distribution is $(\nu+1)(\nu+3)^{-1}T^{-1}$ based on the Fisher information for a t-distribution. An estimate of the rate of missing information due to Y_{mis} alone (the overall population rate of missing information) is

$$\hat{\lambda} = \frac{B + (1 - M^{-1})W_1 + (1 - N^{-1})W_2}{\bar{U} + B + (1 - M^{-1})W_1 + (1 - N^{-1})W_2}. \tag{3.1}$$

This measure does not account for information lost due to a finite number of imputations. Similarly, if Y_{mis}^{A} were observed, then the between-nest variance would be zero. Then the estimated population rate of information due to Y_{mis}^{B} and Y_{mis}^{C} given Y_{mis}^{A} is

$$\hat{\lambda}^{B,C|A} = \frac{W_1 + (1 - N^{-1})W_2}{\bar{U} + W_1 + (1 - N^{-1})W_2}. \tag{3.2}$$

Additionally, if Y_{mis}^{A} and Y_{mis}^{B} were observed, then both the between-nest and within-nest variance would vanish so the estimated population rate of missing information due to Y_{mis}^{C} given Y_{mis}^{A} and Y_{mis}^{B} would be

$$\hat{\lambda}^{C|A,B} = \frac{W_2}{\bar{U} + W_2}. \tag{3.3}$$

The difference $\hat{\lambda}^{A} = \hat{\lambda} - \hat{\lambda}^{B,C|A}$ represents the amount by which the rate of missing information would drop if Y_{mis}^{A} were known. The difference $\hat{\lambda}^{B|A} = \hat{\lambda}^{B,C|A} - \hat{\lambda}^{C|A,B}$ represents the amount by which the rate of missing information would drop if Y_{mis}^{A} and Y_{mis}^{B} were known.

3.2 Asymptotic Distribution of Missing Information Rates

Based on the ANOVA table presented in Table 1 and from analysis of variance literature, it is known that the mean squares are independent chi-squared random variables such that

$$\left(\frac{\sigma^2_{W_2} + N\sigma^2_{W_1} + MN\sigma^2_B}{L-1}\right)^{-1} MNB \sim \chi^2_{L-1}$$
$$\left(\frac{\sigma^2_{W_2} + N\sigma^2_{W_1}}{L(M-1)}\right)^{-1} NW_1 \sim \chi^2_{L(M-1)}$$
$$\left(\frac{\sigma^2_{W_2}}{LM(N-1)}\right)^{-1} W_2 \sim \chi^2_{LM(N-1)}.$$

Based on these distributions, the expectations and variances of B, W_1, and W_2 are as follows:

$$E(B) = \frac{\sigma^2_{W_2} + N\sigma^2_{W_1} + MN\sigma^2_B}{MN}$$
$$V(B) = \frac{2(\sigma^2_{W_2} + N\sigma^2_{W_1} + MN\sigma^2_B)^2}{(L-1)M^2N^2}$$
$$E(W_1) = \frac{\sigma^2_{W_2} + N\sigma^2_{W_1}}{N}$$
$$V(W_1) = \frac{2(\sigma^2_{W_2} + N\sigma^2_{W_1})^2}{L(M-1)N^2}$$
$$E(W_2) = \sigma^2_{W_2}$$
$$V(W_2) = \frac{2(\sigma^2_{W_2})^2}{LM(N-1)}.$$

A first-order approximation to the joint distribution of B, W_1, and W_2 is

$$\sqrt{L}\begin{pmatrix} V_1 & 0 & 0 \\ 0 & V_2 & 0 \\ 0 & 0 & V_3 \end{pmatrix}^{-1}\left(\begin{pmatrix} B \\ W_1 \\ W_2 \end{pmatrix} - E\begin{pmatrix} B \\ W_1 \\ W_2 \end{pmatrix}\right) \sim N(0, I)$$

where $V_1 = 2(\sigma^2_{W_2} + N\sigma^2_{W_1} + MN\sigma^2_B)^2/(M^2N^2)$, $V_2 = 2(\sigma^2_{W_2} + N\sigma^2_{W_1})^2/((M-1)N^2)$, and $V_3 = 2(\sigma^2_{W_2})^2/(M(N-1))$. The limiting distribution for the rates of missing information can then be obtained using the multivariate delta method. Through the multivariate delta method, it is possible to find the limiting distributions of λ, $\lambda^{B,C|A}$, and $\lambda^{C|A,B}$. Based on this multivariate normal distribution, it is straightforward to find the distributions of λ^A and $\lambda^{B|A}$ using a linear combination of λ, $\lambda^{B,C|A}$, and $\lambda^{C|A,B}$. λ, λ^A, $\lambda^{B|A}$, and $\lambda^{C|A,B}$ are the main parameters of interest. It follows that

$$\sqrt{L}\Sigma^{-1}\left(\begin{pmatrix} \lambda \\ \lambda^{B,C|A} \\ \lambda^{C|A,B} \end{pmatrix} - \begin{pmatrix} \hat{\lambda} \\ \hat{\lambda}^{B,C|A} \\ \hat{\lambda}^{C|A,B} \end{pmatrix}\right) \sim N(0, I)$$

where

$$
\begin{aligned}
\Sigma_{11} =& \frac{M^2N^2\bar{U}^2(N^2V_2 - 2MN^2V_2 + M^2(V_3 - 2NV_3 + N^2(V_1 + V_2 + V_3)))}{(MNB - NW_1 + M(-W_2 + N(\bar{U} + W_1 + W_2)))^4} \\
\Sigma_{12} =& \frac{MN^2\bar{U}^2(MV_3 - 2MNV_3 + N^2((M-1)V_2 + MV_3))}{N^4(\bar{U} + W_1 + (1 - N^{-1})W_2)^2M^2(\bar{U} + B + (1 - M^{-1})W_1 + (1 - N^{-1})W_2)^2} \\
\Sigma_{13} =& \frac{M^2(N-1)N\bar{U}^2V_3}{(\bar{U} + W_2)^2(MNB - NW_1 + M(-W_2 + N(\bar{U} + W_1 + W_2)))^2} \\
\Sigma_{21} =& \Sigma_{12} \\
\Sigma_{22} =& \frac{N^2\bar{U}^2(V_3 - 2NV_3 + N^2(V_2 + V_3))}{(W_2 - N(\bar{U} + W_1 + W_2))^4} \\
\Sigma_{23} =& \frac{(N-1)N\bar{U}^2V_3}{(\bar{U} + W_2)^2(W_2 - N(\bar{U} + W_1 + W_2))^2} \\
\Sigma_{31} =& \Sigma_{13} \\
\Sigma_{32} =& \Sigma_{23} \\
\Sigma_{33} =& \frac{\bar{U}^2V_3}{(\bar{U} + W_2)^4}.
\end{aligned}
$$

The covariance matrix, Σ, can be rewritten as a function of λ, $\lambda^{B,C|A}$, and $\lambda^{C|A,B}$ as follows:

$$
\begin{aligned}
\Sigma_{11} =& 2(1-\lambda)^4 \left[\left(\frac{\lambda}{1-\lambda} - \frac{\lambda^{B,C|A}}{1-\lambda^{B,C|A}} \frac{M-1}{M} - \frac{\lambda^{C|A,B}}{1-\lambda^{C|A,B}} \frac{N-1}{MN} \right)^2 \right. \\
& + \frac{M-1}{M^2} \left(\frac{\lambda^{B,C|A}}{1-\lambda^{B,C|A}} - \frac{\lambda^{C|A,B}}{1-\lambda^{C|A,B}} \frac{N-1}{N} \right)^2 \\
& \left. + \frac{N-1}{MN^2} \left(\frac{\lambda^{C|A,B}}{1-\lambda^{C|A,B}} \right)^2 \right] \\
\Sigma_{12} =& \frac{2(1-\lambda)^2(1-\lambda^{B,C|A})^2}{M} \left[\left(\frac{\lambda^{B,C|A}}{1-\lambda^{B,C|A}} - \frac{\lambda^{C|A,B}}{1-\lambda^{C|A,B}} \frac{N-1}{N} \right)^2 \right. \\
& \left. + \frac{N-1}{N^2} \left(\frac{\lambda^{C|A,B}}{1-\lambda^{C|A,B}} \right)^2 \right] \\
\Sigma_{13} =& \frac{2(1-\lambda)^2(\lambda^{C|A,B})^2}{MN} \\
\Sigma_{21} =& \Sigma_{12} \\
\Sigma_{22} =& 2(1-\lambda^{B,C|A})^4 \left[\frac{1}{M-1} \left(\frac{\lambda^{B,C|A}}{1-\lambda^{B,C|A}} - \frac{\lambda^{C|A,B}}{1-\lambda^{C|A,B}} \frac{N-1}{N} \right)^2 \right. \\
& \left. + \frac{N-1}{MN^2} \left(\frac{\lambda^{C|A,B}}{1-\lambda^{C|A,B}} \right)^2 \right] \\
\Sigma_{23} =& \frac{2(1-\lambda^{B,C|A})^2(\lambda^{C|A,B})^2}{MN} \\
\Sigma_{31} =& \Sigma_{13} \\
\Sigma_{32} =& \Sigma_{23} \\
\Sigma_{33} =& \frac{2(\lambda^{C|A,B})^2(1-\lambda^{C|A,B})^2}{M(N-1)}.
\end{aligned}
$$

Confidence intervals may be obtained using this distribution and knowing the values of L, M, N, and estimated rates of missing information for each type of missing value.

3.3 Simulations

When the rates of missing information are of interest to the researcher, more imputations are needed than would be required for stable estimates for a parameter. Estimates for rates of missing information have been shown to be noisy for small numbers of imputations (Schafer, 1997). These simulations attempt to ascertain how many imputations are necessary for stable estimates of the rates of missing information. This section includes simulations which compare theoretical values and simulated values using a finite number of imputations. Estimates for the rates of missing information and the associated standard errors are included.

3.3.1 Set-Up

Let Y consist of 100 independent draws from a normal distribution. Missing values were imposed with a MCAR structure in that the assigned rates of missing information at each stage designated the number of values to be deleted. The MCAR structure was imposed so that the percentage of missing values corresponds to the rates of missing information at each stage. For instance, $\lambda^{A} = .25$, $\lambda^{B|A} = .20$, and $\lambda^{C|A,B} = .05$ correspond to randomly selecting and deleting 25 values at the first stage, 20 values at the second stage, and 5 values at the third stage. The missing values were then imputed, varying the number of imputations at each stage, using the norm package in R (Schafer & Novo, 2013). The number of imputations were chosen to represent small, medium, and large numbers of imputations. The process was repeated 1000 times and analyzed for estimates for the rates of missing information, the corresponding standard errors (as presented in the previous section), and coverage. The nominal coverage is 95%.

The three combinations of rates of missing information include:

1. $\lambda = .50$, $\lambda^A = .25$, $\lambda^{B|A} = .20$, and $\lambda^{C|A,B} = .05$ (Table 5).

2. $\lambda = .25$, $\lambda^A = .10$, $\lambda^{B|A} = .10$, and $\lambda^{C|A,B} = .05$ (Table 6).

3. $\lambda = .15$, $\lambda^A = .05$, $\lambda^{B|A} = .05$, and $\lambda^{C|A,B} = .05$ (Table 7).

3.3.2 Summary of Results

The results are presented in Tables 5-7. The estimates for the rates of missing information converge to the true values and stabilize as the number of imputations increases. Additionally, the coverage is consistently above 84.9%, showing nominal increases as the number of imputations increases. The simulated estimates of the standard errors are close to the theoretical values of the standard errors (within 5% bias consistently), implying that the theoretical estimates presented are good estimates of the true standard errors.

(L,M,N)	$\hat{\lambda}$	$\hat{\lambda}^{A}$	$\hat{\lambda}^{B\|A}$	$\hat{\lambda}^{C\|A,B}$	$\sqrt{V(\hat{\lambda})}$	$\sqrt{V(\hat{\lambda}^{A})}$	$\sqrt{V(\hat{\lambda}^{B\|A})}$	$\sqrt{V(\hat{\lambda}^{C\|A,B})}$
	(Coverage)				**Theoretical Value**			
(10,2,2)	0.480	0.236	0.194	0.049	0.093	0.108	0.075	0.015
	(0.894)	(0.894)	(0.882)	(0.907)	**0.093**	**0.110**	**0.078**	**0.015**
(10,5,2)	0.487	0.239	0.197	0.050	0.084	0.087	0.040	0.010
	(0.892)	(0.877)	(0.930)	(0.939)	**0.083**	**0.085**	**0.039**	**0.010**
(10,5,5)	0.491	0.242	0.198	0.050	0.085	0.088	0.036	0.005
	(0.892)	(0.881)	(0.923)	(0.917)	**0.082**	**0.084**	**0.037**	**0.005**
(50,2,2)	0.501	0.251	0.200	0.050	0.043	0.050	0.035	0.007
	(0.933)	(0.933)	(0.937)	(0.923)	**0.042**	**0.049**	**0.035**	**0.007**
(50,5,2)	0.499	0.249	0.199	0.050	0.039	0.041	0.019	0.004
	(0.921)	(0.920)	(0.921)	(0.943)	**0.037**	**0.038**	**0.018**	**0.004**
(50,5,5)	0.500	0.249	0.201	0.050	0.040	0.040	0.018	0.003
	(0.910)	(0.922)	(0.919)	(0.894)	**0.037**	**0.038**	**0.016**	**0.002**
(100,2,2)	0.501	0.250	0.201	0.050	0.032	0.036	0.025	0.005
	(0.911)	(0.933)	(0.944)	(0.931)	**0.030**	**0.035**	**0.025**	**0.005**
(100,5,2)	0.501	0.250	0.200	0.050	0.029	0.029	0.014	0.003
	(0.930)	(0.921)	(0.917)	(0.926)	**0.026**	**0.027**	**0.012**	**0.003**
(100,5,5)	0.501	0.250	0.201	0.050	0.028	0.027	0.013	0.002
	(0.923)	(0.942)	(0.925)	(0.869)	**0.026**	**0.027**	**0.012**	**0.002**

Table 5: Estimates of rates of missing information and associated standard errors when $\lambda = .50$, $\lambda^{A} = .25$, $\lambda^{B|A} = .20$, and $\lambda^{C|A,B} = .05$. Coverage values are displayed in parentheses and theoretical values for the standard errors are displayed in bold. The number of imputations at each stage are summarized in the ordered triple, (L,M,N)

(L, M, N)	$\hat{\lambda}$	$\hat{\lambda}^{A}$	$\hat{\lambda}^{B\|A}$	$\hat{\lambda}^{C\|A,B}$	$\sqrt{V(\hat{\lambda})}$	$\sqrt{V(\hat{\lambda}^{A})}$	$\sqrt{V(\hat{\lambda}^{B\|A})}$	$\sqrt{V(\hat{\lambda}^{C\|A,B})}$
	(Coverage)				**Theoretical Value**			
(10,2,2)	0.248	0.098	0.100	0.050	0.064	0.067	0.048	0.015
	(0.898)	(0.896)	(0.904)	(0.897)	**0.062**	**0.065**	**0.049**	**0.015**
(10,5,2)	0.248	0.097	0.100	0.051	0.051	0.049	0.025	0.010
	(0.889)	(0.855)	(0.928)	(0.940)	**0.049**	**0.048**	**0.025**	**0.010**
(10,5,5)	0.249	0.098	0.101	0.050	0.052	0.049	0.023	0.005
	(0.875)	(0.849)	(0.931)	(0.935)	**0.048**	**0.047**	**0.022**	**0.005**
(50,2,2)	0.254	0.102	0.102	0.050	0.028	0.032	0.022	0.006
	(0.943)	(0.915)	(0.944)	(0.953)	**0.028**	**0.029**	**0.022**	**0.007**
(50,5,2)	0.253	0.102	0.099	0.051	0.024	0.023	0.012	0.004
	(0.939)	(0.939)	(0.923)	(0.950)	**0.022**	**0.021**	**0.011**	**0.004**
(50,5,5)	0.250	0.100	0.099	0.050	0.022	0.020	0.010	0.002
	(0.931)	(0.940)	(0.941)	(0.960)	**0.022**	**0.021**	**0.010**	**0.002**
(100,2,2)	0.254	0.102	0.102	0.050	0.020	0.023	0.017	0.005
	(0.949)	(0.929)	(0.941)	(0.927)	**0.019**	**0.021**	**0.016**	**0.005**
(100,5,2)	0.254	0.103	0.100	0.051	0.017	0.016	0.009	0.003
	(0.937)	(0.940)	(0.922)	(0.962)	**0.016**	**0.015**	**0.008**	**0.003**
(100,5,5)	0.249	0.100	0.099	0.050	0.017	0.016	0.007	0.002
	(0.925)	(0.927)	(0.937)	(0.920)	**0.015**	**0.015**	**0.007**	**0.002**

Table 6: Estimates of rates of missing information and associated standard errors when $\lambda = .25$, $\lambda^{A} = .10$, $\lambda^{B|A} = .10$, and $\lambda^{C|A,B} = .05$. Coverage values are displayed in parentheses and theoretical values for the standard errors are displayed in bold. The number of imputations at each stage are summarized in the ordered triple, (L, M, N)

(L, M, N)	$\hat{\lambda}$	$\hat{\lambda}^{A}$	$\hat{\lambda}^{B\|A}$	$\hat{\lambda}^{C\|A,B}$	$\sqrt{V(\hat{\lambda})}$	$\sqrt{V(\hat{\lambda}^{A})}$	$\sqrt{V(\hat{\lambda}^{B\|A})}$	$\sqrt{V(\hat{\lambda}^{C\|A,B})}$
		(Coverage)				**Theoretical Value**		
(10,2,2)	0.151	0.051	0.051	0.050	0.040	0.038	0.031	0.015
	(0.907)	(0.921)	(0.905)	(0.882)	**0.038**	**0.039**	**0.032**	**0.015**
(10,5,2)	0.149	0.049	0.049	0.051	0.029	0.027	0.016	0.010
	(0.909)	(0.861)	(0.937)	(0.944)	**0.029**	**0.027**	**0.016**	**0.010**
(10,5,5)	0.152	0.050	0.051	0.051	0.028	0.027	0.013	0.005
	(0.908)	(0.862)	(0.932)	(0.947)	**0.027**	**0.026**	**0.013**	**0.005**
(50,2,2)	0.152	0.050	0.051	0.050	0.018	0.018	0.015	0.007
	(0.939)	(0.938)	(0.942)	(0.927)	**0.017**	**0.017**	**0.014**	**0.007**
(50,5,2)	0.151	0.050	0.050	0.051	0.014	0.013	0.007	0.005
	(0.942)	(0.920)	(0.952)	(0.932)	**0.013**	**0.012**	**0.007**	**0.004**
(50,5,5)	0.152	0.051	0.051	0.051	0.013	0.012	0.006	0.002
	(0.944)	(0.932)	(0.945)	(0.933)	**0.012**	**0.012**	**0.006**	**0.002**
(100,2,2)	0.152	0.051	0.051	0.050	0.012	0.013	0.010	0.005
	(0.945)	(0.937)	(0.951)	(0.923)	**0.012**	**0.012**	**0.010**	**0.005**
(100,5,2)	0.152	0.051	0.050	0.051	0.010	0.009	0.006	0.003
	(0.932)	(0.931)	(0.934)	(0.923)	**0.009**	**0.009**	**0.005**	**0.003**
(100,5,5)	0.152	0.051	0.051	0.051	0.009	0.008	0.004	0.002
	(0.945)	(0.940)	(0.942)	(0.872)	**0.009**	**0.008**	**0.004**	**0.002**

Table 7: Estimates of rates of missing information and associated standard errors when $\lambda = .15$, $\lambda^{A} = .05$, $\lambda^{B|A} = .05$, and $\lambda^{C|A,B} = .05$. Coverage values are displayed in parentheses and theoretical values for the standard errors are displayed in bold. The number of imputations at each stage are summarized in the ordered triple, (L, M, N)

Chapter 4

Ignorability

Chapter 1 describes ignorability as being the weakest set of conditions under which the distribution of the missing data process does not need to be modeled in Bayesian or likelihood-based inferences. These two conditions include that the missingness mechanism is MAR and that the parameter of interest, θ, and the parameter of the missingness, ϕ^+, are distinct.

The definitions become more complicated when the missing data are partitioned into two or more types. Harel (2009) and Harel & Schafer (2009) explore the situation where there are two types of missing values. Challenges in ignorability include the possibility of one type of missing value being dependent on the other type of missing value. Another challenge is the situation where a missing datum may be at risk of being missing for both of the reasons. As a result, Harel (2009) extends the definitions of MAR and ignorability presented by Rubin (1976). The purpose of this chapter is to further extend those results to include three types of missing values. Throughout this chapter, M^+ will refer to the missingness matrix as described in Chapters 1 and 2. In standard multiple imputation, M^+ can take on values of 0 or 1. In two-stage multiple imputation, M^+ can take on values of 0, 1, or 2. In three-stage multiple imputation, M^+ can take on values of 0, 1, 2, or 3.

4.1 Standard Ignorability

Rubin (1976) defines missing at random (MAR) as follows:

Definition 4.1. *Let M^+ be a set of indicator random variables that separate the complete data Y_{com} into (Y_{obs}, Y_{mis}) and let ϕ^+ be the parameter of the conditional distribution of M^+ given Y_{com}. Data are considered to be MAR if*

$$P(M^+|Y_{obs}, Y_{mis}, \phi^+) = P(M^+|Y_{obs}, \phi^+) \tag{4.1}$$

for all possible values of ϕ^+, at the realized values of M^+ and Y_{obs}.

Conceptually, this equates to the distribution of missingness being functionally independent of the missing part of the data. A missing data mechanism is said to be ignorable if MAR holds and θ and ϕ^+ are distinct (Little & Rubin, 2002). Under ignorability, the model for M^+ given Y_{com} is irrelevant for likelihood-based or Bayesian inferences about θ. That is, the joint likelihood for (Y_{obs}, M^+) given (θ, ϕ^+) is

$$\begin{aligned}
P(Y_{obs}, M^+|\theta, \phi^+) &= \int P(Y_{com}, M^+|\theta, \phi^+) dY_{mis} \\
&= \int P(Y_{obs}, Y_{mis}|\theta) P(M^+|Y_{obs}, Y_{mis}, \phi^+) dY_{mis} \\
&= P(M^+|Y_{obs}, \phi^+) \int P(Y_{obs}, Y_{mis}|\theta) dY_{mis} \\
&= P(M^+|Y_{obs}, \phi^+) P(Y_{obs}|\theta) \\
&\propto L(\phi^+|M^+, Y_{obs}) L(\theta|Y_{obs})
\end{aligned}$$

where $L(\phi^+|M^+, Y_{obs})$ is a likelihood function that does not involve θ and $L(\theta|Y_{obs})$ is a likelihood function that does not involve ϕ^+.

4.2 Two-Stage Ignorability

Harel (2009) and Harel & Schafer (2009) describe extended missingness when there are two types of missing values. Let the complete data be partitioned into $(Y_{obs}, Y_{mis}^A, Y_{mis}^B)$ and let M^+ be the extended missing data matrix. Let ϕ^+ be the parameter of the missing data process.

Definition 4.2. *The missing data $Y_{mis} = (Y_{mis}^A, Y_{mis}^B)$ are said to be MAR$^+$ if*

$$P(M^+|Y_{obs}, Y_{mis}, \phi^+) = P(M^+|Y_{obs}, \phi^+) \tag{4.2}$$

for all possible ϕ^+, at the realized values of M^+ and Y_{obs}.

This condition is stronger than MAR because it assumes that the process that subdivides Y_{mis} does not depend on Y_{mis} either.

Harel (2009) concludes that if θ and ϕ^+ are distinct and MAR$^+$ holds, then the information contained in M^+ can be ignored when making likelihood or Bayesian inferences for θ. The implication here is that M^+ may be ignored when imputing both Y^A_{mis} and Y^B_{mis}.

Some weaker conditions can be imposed such that some aspects of M^+ can be ignored in one or both stages.

Definition 4.3. *The missing data Y^B_{mis} in a two-stage setting are said to be conditionally missing at random (CMAR$^+$) if*

$$P(M^+|Y_{obs}, Y_{mis}, \phi^+) = P(M^+|Y_{obs}, Y^A_{mis}, \phi^+) \tag{4.3}$$

for all possible ϕ^+, with M^+ and Y_{obs} fixed at their realized values.

If distinctness and CMAR$^+$ hold, then the information in M^+ can be ignored at the second stage of imputation.

This process can be generalized by the idea of latently missing at random (LMAR) presented by Harel & Schafer (2009):

Definition 4.4. *Let $h(Y_{mis})$ denote a coarsened summary or many-to-one function of the missing values. Missing values are said to be latently missing at random given $h(Y_{mis})$*

if

$$P(M^+|Y_{obs}, Y_{mis}, \phi^+) = P(M^+|Y_{obs}, h(Y_{mis}), \phi^+)$$

for all possible ϕ^+, with M^+ and Y_{obs} fixed at their realized values.

4.3 Three-Stage Ignorability

When three types of missing values are present, the data can be partitioned as $Y_{com} = (Y_{obs}, Y^A_{mis}, Y^B_{mis}, Y^C_{mis})$. M^+ remains the set of indicators partitioning the complete data. The purpose of describing extended missingness is not to encourage modeling of M^+ but instead to illustrate conditions under which the modeling of M^+ is unnecessary.

4.3.1 Extended Ignorability

Definition 4.2 for the extended ignorability still holds with three types of missing values. The main idea of the definition is that the process that subdivides Y_{mis} is not dependent on Y_{mis} so the definition is relevant whether the missing data are being divided into two parts, three parts, or more. For completeness, Definition 4.2 can be rewritten as follows:

Definition 4.5. *The missing data Y_{mis}, regardless of the number of partitions, are said to be MAR$^+$ if*

$$P(M^+|Y_{obs}, Y_{mis}, \phi^+) = P(M^+|Y_{obs}, \phi^+)$$

for all possible ϕ^+, at the realized values of M^+ and Y_{obs}.

Definition 4.2 is simply a special case of Definition 4.5 with two partitions and three-stage ignorability is a special case with three partitions.

Result 4.6. *If θ and ϕ^+ are distinct and MAR^+ holds, then we can ignore the information contained in M^+ when making likelihood or Bayesian inferences about θ.*

The proof of this result is identical to that of Rubin's.

Result 4.6 shows that M^+ can be ignored at all imputation stages as indicated by the following result:

Result 4.7. *If θ and ϕ^+ are distinct and MAR^+ holds, then we can ignore the information contained in M^+ in all stages of imputation (2.2), (2.3), (2.4), so that*

$$P(Y^A_{mis}|Y_{obs}, M^+) = P(Y^A_{mis}|Y_{obs})$$

and

$$P(Y^B_{mis}|Y_{obs}, Y^A_{mis}, M^+) = P(Y^B_{mis}|Y_{obs}, Y^A_{mis})$$

and

$$P(Y^C_{mis}|Y_{obs}, Y^A_{mis}, Y^B_{mis}, M^+) = P(Y^C_{mis}|Y_{obs}, Y^A_{mis}, Y^B_{mis}).$$

Proof. The predictive distribution (2.2) for Y^A_{mis} (under distinctness) is

$$
\begin{aligned}
P(Y^A_{mis}|Y_{obs}, M^+) &= \int\int\int\int P(\theta, \phi^+, Y^A_{mis}, Y^B_{mis}, Y^C_{mis}|Y_{obs}, M^+) d\phi^+ d\theta dY^B_{mis} dY^C_{mis} \\
&= \frac{1}{P(Y_{obs}, M^+)} \int\int\int\int P(\theta) P(Y_{com}|\theta) \\
&\quad \times P(\phi^+) P(M^+|Y_{com}, \phi^+) d\phi^+ d\theta dY^B_{mis} dY^C_{mis}.
\end{aligned}
$$

Under MAR$^+$, this becomes

$$
\begin{aligned}
P(Y^A_{mis}|Y_{obs}, M^+) &= \frac{1}{P(Y_{obs}, M^+)} \int\int\int\int P(\theta) P(Y_{com}|\theta) \\
&\quad \times P(\phi^+) P(M^+|Y_{com}, \phi^+) d\phi^+ d\theta dY^B_{mis} dY^C_{mis} \\
&= \frac{P(M^+|Y_{obs})}{P(Y_{obs}, M^+)} \int\int P(Y_{obs}, Y^A_{mis}, Y^B_{mis}, Y^C_{mis}) dY^B_{mis} dY^C_{mis} \\
&= \frac{P(M^+|Y_{obs})}{P(Y_{obs}, M^+)} P(Y_{obs}, Y^A_{mis}) \\
&= P(Y^A_{mis}|Y_{obs}).
\end{aligned}
$$

Similarly, the predictive distribution (2.3) for Y^B_{mis} (under distinctness) is

$$
\begin{aligned}
P(Y^B_{mis}|Y_{obs}, Y^A_{mis}, M^+) &= \int\int\int P(\theta, \phi^+, Y^A_{mis}, Y^B_{mis}, Y^C_{mis}|Y_{obs}, M^+) d\phi^+ d\theta dY^C_{mis} \\
&= \frac{1}{P(Y_{obs}, Y^A_{mis}, M^+)} \int\int\int P(\theta) P(Y_{com}|\theta) \\
&\quad \times P(\phi^+) P(M^+|Y_{com}, \phi^+) d\phi^+ d\theta dY^C_{mis}.
\end{aligned}
$$

Under MAR$^+$, this becomes

$$\begin{aligned}
P(Y^B_{mis}|Y_{obs}, Y^A_{mis}, M^+) &= \frac{1}{P(Y_{obs}, Y^A_{mis}, M^+)} \int\int\int P(\theta)P(Y_{com}|\theta) \\
&\quad \times P(\phi^+)P(M^+|Y_{com}, \phi^+)d\phi^+ d\theta dY^C_{mis} \\
&= \frac{P(M^+|Y_{obs}, Y^A_{mis})}{P(Y_{obs}, Y^A_{mis}, M^+)} \int P(Y_{obs}, Y^A_{mis}, Y^B_{mis}, Y^C_{mis})dY^C_{mis} \\
&= \frac{P(M^+|Y_{obs}, Y^A_{mis})}{P(Y_{obs}, Y^A_{mis}, M^+)} P(Y_{obs}, Y^A_{mis}, Y^B_{mis}) \\
&= P(Y^B_{mis}|Y_{obs}, Y^A_{mis}).
\end{aligned}$$

Finally, the predictive distribution (2.4) for Y^C_{mis} under distinctness and MAR$^+$ is

$$\begin{aligned}
P(Y^C_{mis}|Y_{obs}, Y^A_{mis}, Y^B_{mis}, M^+) &= \int\int P(\theta, \phi^+, Y^A_{mis}, Y^B_{mis}, Y^C_{mis}|Y_{obs}, M^+)d\phi^+ d\theta \\
&= \frac{1}{P(Y_{obs}, Y^A_{mis}, Y^B_{mis}, M^+)} \int\int P(\theta)P(Y_{com}|\theta) \\
&\quad \times P(\phi^+)P(M^+|Y_{com}, \phi^+)d\phi^+ d\theta \\
&= \frac{P(M^+|Y_{obs}, Y^A_{mis}, Y^B_{mis})}{P(Y_{obs}, Y^A_{mis}, Y^B_{mis}, M^+)} P(Y_{obs}, Y^A_{mis}, Y^B_{mis}, Y^C_{mis}) \\
&= P(Y^C_{mis}|Y_{obs}, Y^A_{mis}, Y^B_{mis}).
\end{aligned}$$

$\square$

4.3.2 Conditional Extended Ignorability

It is desirable to define weaker conditions under which M^+ can be ignored in one or more stages. This is similar to the CMAR$^+$ definition presented by Harel (2009). The first simplification occurs if M^+ is dependent on Y^A_{mis} but not on Y^B_{mis} or Y^C_{mis} and the second simplification occurs if M^+ is dependent on Y^A_{mis} and Y^B_{mis} but not on Y^C_{mis}.

Definition 4.8. *The missing data (Y^B_{mis}, Y^C_{mis}) in a three-stage setting are said to be CMAR^2 if*

$$P(M^+|Y_{obs}, Y_{mis}, \phi^+) = P(M^+|Y_{obs}, Y^A_{mis}, \phi^+)$$

for all possible ϕ^+, at the realized values of M^+ and Y_{obs}.

Result 4.9. *If θ and ϕ^+ are distinct and CMAR^2 holds, then we can ignore the information contained in M^+ in the second and third stages of imputation, so that*

$$P(Y^B_{mis}|Y_{obs}, Y^A_{mis}, M^+) = P(Y^B_{mis}|Y_{obs}, Y^A_{mis}) \text{ and}$$
$$P(Y^C_{mis}|Y_{obs}, Y^A_{mis}, Y^B_{mis}, M^+) = P(Y^C_{mis}|Y_{obs}, Y^A_{mis}, Y^B_{mis}).$$

Definition 4.10. *The missing data Y^C_{mis} in a three-stage setting are CMAR^3 if*

$$P(M^+|Y_{obs}, Y_{mis}, \phi^+) = P(M^+|Y_{obs}, Y^A_{mis}, Y^B_{mis}, \phi^+)$$

for all possible ϕ^+, at the realized values of M^+ and Y_{obs}.

Result 4.11. *If θ and ϕ^+ are distinct and CMAR^3 holds, then we can ignore the information contained in M^+ in the third stage of imputation, so that*

$$P(Y_{mis}^C | Y_{obs}, Y_{mis}^A, Y_{mis}^B, M^+) = P(Y_{mis}^C | Y_{obs}, Y_{mis}^A, Y_{mis}^B).$$

All of the situations described thus far partition the missing data into multiple parts. One can also consider factoring M^+ into multiple submodels. This addresses the notion of partially missing at random presented in Harel & Schafer (2009).

Definition 4.12. *Let $g(M^+)$ denote a coarsened summary or many-to-one function of M^+. Suppose we factorize the missing data mechanism as*

$$P(M^+ | Y_{com}, \phi^+) = P(g(M^+) | Y_{com}, \gamma) P(M^+ | Y_{com}, \delta)$$

where $\phi^+ = (\gamma, \delta)$. Missing data are said to be partially missing at random given $g(M^+)$ if

$$P(M^+ | Y_{com}, g(M^+), \delta) = P(M^+ | Y_{obs}, g(M^+), \delta)$$

for all possible δ, with M^+ and Y_{obs} fixed at their realized values.

4.4 Simulations

In order to assess three-stage multiple imputation in the presence of nonignorable missingness, four simulation scenarios were run. The first simulation scenario has one type of missing value which is nonignorable and two types which are ignorable. The second simulation scenario has two types of missing values which are nonignorable and one type which is ignorable. The third simulation scenario also has two types of missing values which are nonignorable and one type which is ignorable but imputes them in a different order to see if the order of imputation has an impact on the analysis. The final simulation scenario has one type of missing value which is nonignorable and two types which are ignorable but the ignorable missingness is in the covariates instead of the response variable.

Additionally, a sensitivity analysis is performed to determine the impact of misspecifying k, the amount by which the nonignorable imputed values are perturbed, as described in Equation 4.4.

4.4.1 One Type of Nonignorable Missing Value

The first simulation scenario considers the situation where only one type of missing value is nonignorable and two types of missing values are ignorable.

Simulation Set-Up

Data for the simulations are generated through the following step. X_1 is generated from a $N(50, 100)$ and X_2 is generated from $N(20, 100)$. 100 values of X_1 and X_2 are randomly drawn. Y is created by taking $Y = 2X_1 + 3X_2$ so that Y has a known mean of 160. Then, Y_{com} is the data which includes Y and X_1.

A missingness structure is imposed as follows:

1. The first type of missing value is created with an MCAR structure. That is, a prespecified percentage of the values in Y are randomly deleted. Let $MCAR\%$ denote the percentage of missing values due to the type of missingness which is MCAR.

2. The second type of missing value is created under an MAR structure. Let $MAR\%$ denote the percentage of missing values due to the type of missingness which is MAR. Values in Y are removed if they are below the bottom $MAR\%$ percentile of X_1.

3. The third type of missing value is created under an MNAR structure. Let the percentage of missing values due to the first type of missingness be denoted $MNAR\%$. Values in Y are removed if they are above the top $MNAR\%$ percentile of Y.

The percent breakdown for each of the three types of missing values for each simulation scenario is presented in Table 8.

Table 8: Breakdown of percentages of missing values by each type (one nonignorable)

Total Percentage of Missing Values	$MCAR\%$	$MAR\%$	$MNAR\%$
75	25	25	25
50	15	15	20
40	15	5	20
15	5	5	5

The missing values are imputed using the norm package in R (Schafer & Novo, 2013) with varying numbers of imputations at each stage denoted by the ordered triple (L, M, N). The order of imputation is $MNAR\%$, $MAR\%$, $MCAR\%$. Additionally, to account for the nonignorable missingness, the imputed values for the nonignorable type are perturbed as described by Rubin (1987). Rubin proposes a few methods for perturbing the imputed values of Y but the simplest is the transformation

$$(\text{nonignorable imputed } Y_i) = k \times (\text{ignorable imputed } Y_i) \tag{4.4}$$

where k is an arbitrary value selected by the imputer based on prior knowledge of where the missing values are expected to lie. For this simulation, $k = 1.2$ is used to reflect that the nonignorable missing values in Y are based on the upper percentile of Y so the values should be 20% higher than the ignorable counterparts. The sensitivity of this assumption is discussed in Section 4.4.4.

The parameter of interest, Q, is the mean of Y. Data are generated, imputed, and analyzed 1000 times for percent bias, MSE, and coverage. The nominal coverage is 95%.

Results

The results of this simulation are presented in Table 9. The percent bias is under 2% and the coverage is between 89.6% and 94.0%. The MSE increases as the percentage of missingness increases. There does not appear to be any discernible pattern to imply that increasing the number of imputations improves percent bias, MSE, or coverage. In some cases, the larger number of imputations is nominally detrimental to the coverage.

Percent Missing	(L,M,N)	Percent Bias (%)	MSE	Coverage
75	(10,2,2)	-0.165	43.004	0.932
	(10,5,2)	-0.227	42.575	0.929
	(10,5,5)	-0.237	42.481	0.925
	(50,2,2)	0.328	41.238	0.933
	(50,5,2)	0.267	40.870	0.936
	(50,5,5)	0.262	40.833	0.935
	(100,2,2)	0.328	40.694	0.938
	(100,5,2)	0.285	40.590	0.935
	(100,5,5)	0.276	40.572	0.935
50	(10,2,2)	-1.270	25.016	0.935
	(10,5,2)	-1.364	25.403	0.930
	(10,5,5)	-1.330	25.168	0.932
	(50,2,2)	-1.339	25.003	0.925
	(50,5,2)	-1.316	24.850	0.925
	(50,5,5)	-1.284	24.639	0.926
	(100,2,2)	-1.468	26.050	0.919
	(100,5,2)	-1.458	25.980	0.916
	(100,5,5)	-1.443	25.869	0.916
40	(10,2,2)	-0.876	20.657	0.923
	(10,5,2)	-1.083	21.619	0.930
	(10,5,5)	-1.067	21.551	0.932
	(50,2,2)	-1.613	25.150	0.896
	(50,5,2)	-1.579	24.866	0.900
	(50,5,5)	-1.570	24.796	0.899
	(100,2,2)	-1.560	24.672	0.903
	(100,5,2)	-1.545	24.562	0.903
	(100,5,5)	-1.557	24.657	0.903
15	(10,2,2)	-0.768	16.259	0.928
	(10,5,2)	-0.817	16.507	0.931
	(10,5,5)	-0.854	16.670	0.930
	(50,2,2)	-0.550	15.571	0.938
	(50,5,2)	-0.548	15.571	0.939
	(50,5,5)	-0.542	15.555	0.939
	(100,2,2)	-0.546	15.525	0.940
	(100,5,2)	-0.543	15.525	0.939
	(100,5,5)	-0.543	15.526	0.939

Table 9: Estimates of percent bias, MSE, and coverage varying the percentages of missing values and number of imputations at each stage for one type of nonignorable missing value

4.4.2 Two Types of Nonignorable Missing Values

The second and third simulations involve the situation where two types of missing values are nonignorable and one type is ignorable.

Set-Up

The data for the simulations is generated through the same process as Section 4.4.1. The difference lies in the missingness structure which is imposed as follows:

1. The first type of missing value is created with an MCAR structure. That is, a prespecified percentage of the values in Y are randomly deleted. Let $MCAR\%$ denote the percentage of missing values due to the type of missingness which is MCAR.

2. The second type of missing value is created under an MNAR structure. Let $MNAR_1\%$ denote the percentage of missing values due to the first type of missingness which is MNAR. Values in Y are removed if they are above the top $MNAR_1\%$ percentile of Y. The missingness is dependent on the value of the response so those values are nonignorable.

3. The third type of missing value is created under an MNAR structure. Let the percentage of missing values due to the second type of MNAR missingness be denoted $MNAR_2\%$. Values in Y are removed if they are below the bottom $MNAR_2\%$ percentile of Y.

Table 10: Breakdown of percentages of missing values by each type (two nonignorable)

Total Percentage of Missing Values	$MCAR\%$	$MNAR_2\%$	$MNAR_1\%$
75	25	25	25
50	15	15	20
40	15	5	20
15	5	5	5

The percent breakdown for each of the three types of missing values is presented in Table 10.

Two orders of imputation were simulated and compared. The first order of imputation is $MNAR_1\%$, $MNAR_2\%$, and $MCAR\%$ (denoted Order 1). The second order reverses the two MNAR mechanisms (denoted Order 2). The nonignorable missing values must be imputed first due to the order of integration as described in Definition 4.10. The imputations are perturbed based on the process described in Section 4.4.1. The values that are $MNAR_1\%$ are multiplied by 1.2 based on the belief that those values are 20% higher than the ignorable counterparts and the values that are $MNAR_2\%$ are multiplied by 0.8 based on the belief that those values are 20% lower than the ignorable counterparts. These numbers were chosen based on the knowledge that $MNAR_1\%$ removes values based on the upper percentile of Y and $MNAR_2\%$ removes values based on the lower percentile of Y.

Results

Table 11 compares the percent bias, MSE, and coverage for the two orders of imputation. It shows that, for each of the percentages of missing values, there is higher percent bias and higher MSE for Order 2. The coverage does not appear to have any significant pattern and ranges from 86.3% to 94.9%. The percent bias in all cases is not larger than 2.5%. In terms of increasing the number of imputations, the coverage actually appears to decrease as the number of imputations increase, except in the case of 15% missingness where the estimates of coverage are fairly stable. This implies that a great number of imputations is not needed.

Percent Missing	(L,M,N)	Percent Bias (%)		MSE		Coverage	
		Order 1	Order 2	Order 1	Order 2	Order 1	Order 2
75	(10,2,2)	1.907	-2.178	29.836	32.784	0.932	0.896
	(10,5,2)	1.817	-2.166	28.901	32.733	0.914	0.865
	(10,5,5)	1.734	-2.250	28.109	33.679	0.906	0.849
	(50,2,2)	1.892	-1.948	29.541	30.355	0.907	0.879
	(50,5,2)	1.811	-1.965	28.742	30.570	0.902	0.869
	(50,5,5)	1.769	-2.005	28.347	30.976	0.898	0.865
	(100,2,2)	1.847	-1.978	29.082	30.567	0.900	0.872
	(100,5,2)	1.837	-1.986	28.980	30.707	0.895	0.866
	(100,5,5)	1.821	-2.002	28.823	30.865	0.894	0.863
50	(10,2,2)	0.243	-0.419	17.296	18.278	0.949	0.934
	(10,5,2)	0.230	-0.460	17.274	18.282	0.937	0.938
	(10,5,5)	0.260	-0.426	17.320	18.219	0.938	0.937
	(50,2,2)	0.239	-0.432	17.348	18.055	0.942	0.919
	(50,5,2)	0.343	-0.402	17.516	17.978	0.939	0.921
	(50,5,5)	0.374	-0.394	17.582	17.962	0.935	0.920
	(100,2,2)	0.117	-0.440	17.226	18.057	0.940	0.919
	(100,5,2)	0.158	-0.494	17.273	18.174	0.936	0.918
	(100,5,5)	0.172	-0.493	17.288	18.171	0.937	0.916
40	(10,2,2)	-0.170	-0.804	17.531	19.095	0.916	0.924
	(10,5,2)	-0.349	-0.973	17.693	19.714	0.917	0.905
	(10,5,5)	-0.331	-0.928	17.678	19.511	0.929	0.900
	(50,2,2)	-0.850	-0.859	19.007	19.157	0.906	0.907
	(50,5,2)	-0.832	-0.841	18.908	19.080	0.907	0.912
	(50,5,5)	-0.822	-0.798	18.862	18.903	0.907	0.911
	(100,2,2)	-0.814	-0.864	18.852	19.203	0.906	0.909
	(100,5,2)	-0.808	-0.844	18.819	19.100	0.906	0.911
	(100,5,5)	-0.819	-0.840	18.860	19.081	0.906	0.911
15	(10,2,2)	-0.054	0.112	14.019	14.054	0.925	0.938
	(10,5,2)	-0.066	0.093	14.013	14.078	0.926	0.936
	(10,5,5)	-0.101	0.057	14.023	14.061	0.927	0.938
	(50,2,2)	0.165	0.221	14.105	14.141	0.934	0.934
	(50,5,2)	0.164	0.180	14.101	14.095	0.934	0.934
	(50,5,5)	0.169	0.185	14.107	14.100	0.934	0.933
	(100,2,2)	0.167	0.237	14.087	14.166	0.934	0.937
	(100,5,2)	0.169	0.225	14.087	14.149	0.934	0.937
	(100,5,5)	0.169	0.224	14.088	14.149	0.934	0.936

Table 11: Comparison of estimates of percent bias, MSE, and coverage varying the percentages of missing values and number of imputations changing the order of imputation for two types of nonignorable missingness. Order 1 imputes $MNAR_1\%$, $MNAR_2\%$, then $MCAR\%$ and Order 2 imputes $MNAR_2\%$, $MNAR_1\%$, then $MCAR\%$

4.4.3 Multivariate Missingness

This simulation involves one type of nonignorable missing value in the response and two types of ignorable missing values in the covariates.

Simulation Set-Up

Data for the simulations are generated through the following steps. X_1 is generated from a $N(50, 100)$ and X_2 is generated from $N(20, 100)$. 100 values of X_1 and X_2 are randomly drawn. X_3 is generated from a $N(0, 1)$ to simulate the error term. Y is created by taking $Y = 2X_1 + 3X_2 + X_3$ so that Y has a known mean of 160. Then, Y_{com} is the data which includes Y, X_1, and X_2.

A missingness structure is imposed as follows:

1. The first type of missing value is created with an MCAR structure. That is, a prespecified percentage of the values in X_2 are randomly deleted. Let $MCAR\%$ denote the percentage of missing values due to the type of missingness which is MCAR.

2. The second type of missing value is created under an MAR structure. Let $MAR\%$ denote the percentage of missing values due to the type of missingness which is MAR. Values in X_1 are removed if they are below the bottom $MAR\%$ percentile of X_2.

3. The third type of missing value is created under an MNAR structure. Denote

$MNAR\%$ as the percentage of missing values due to the type of missingness which is MNAR. Values in Y are removed if they are above the top $MNAR\%$ percentile of Y. The missingness here is dependent on the values of the response so those values are MNAR and therefore, nonignorable.

The percent breakdown for each of the three types of missing values for each simulation scenario is presented in Table 8. The missing values are imputed using the norm package in R (Schafer & Novo, 2013) with varying numbers of imputations at each stage denoted by the ordered triple (L, M, N). The order of imputation is $MNAR\%$, $MAR\%$, $MCAR\%$. Additionally, to account for the nonignorable missingness, the imputed values for the nonignorable type are perturbed with k=1.2 indicating that the nonignorable missing values are 20% larger than their ignorable counterparts. Choice of k is discussed further in Section 4.4.4.

The parameter of interest, Q, is the mean of Y. Data are generated, imputed, and analyzed 1000 times for bias, MSE, and coverage. The nominal coverage is 95%.

Results

The results of this simulation are presented in Table 12. The percent bias ranges from 1.3% to 4.9% with the bias increasing as the percentage of missingness increases. The MSE also increases as the percentage of missingness increases. The coverage ranges from 68.8% to 95.1% with the coverage improving as the percentage of missingness decreases. There is little variability in the estimates when the numbers of imputations are changed but this is largely due to the fact that the parameter of interest is the mean of Y so the number of imputations for the second and third type (which are also imputed after the missing values in Y) offer no changes to the overall estimate of the mean of Y.

Percent Missing	(L,M,N)	Percent Bias (%)	MSE	Coverage
75	(10,2,2)	4.900	77.690	0.688
	(10,5,2)	4.900	77.690	0.688
	(10,5,5)	4.900	77.690	0.688
	(50,2,2)	4.770	74.551	0.705
	(50,5,2)	4.770	74.551	0.705
	(50,5,5)	4.770	74.551	0.705
	(100,2,2)	4.805	75.404	0.703
	(100,5,2)	4.805	75.404	0.703
	(100,5,5)	4.805	75.404	0.703
50	(10,2,2)	4.346	63.526	0.755
	(10,5,2)	4.346	63.526	0.755
	(10,5,5)	4.346	63.526	0.755
	(50,2,2)	4.348	63.654	0.751
	(50,5,2)	4.348	63.654	0.751
	(50,5,5)	4.348	63.654	0.751
	(100,2,2)	4.383	64.401	0.747
	(100,5,2)	4.383	64.401	0.747
	(100,5,5)	4.383	64.401	0.747
40	(10,2,2)	4.460	66.037	0.739
	(10,5,2)	4.460	66.037	0.739
	(10,5,5)	4.460	66.037	0.739
	(50,2,2)	4.487	66.711	0.736
	(50,5,2)	4.487	66.711	0.736
	(50,5,5)	4.487	66.711	0.736
	(100,2,2)	4.468	66.266	0.739
	(100,5,2)	4.468	66.266	0.739
	(100,5,5)	4.468	66.266	0.739
15	(10,2,2)	1.365	18.415	0.951
	(10,5,2)	1.365	18.415	0.951
	(10,5,5)	1.365	18.415	0.951
	(50,2,2)	1.374	18.461	0.951
	(50,5,2)	1.374	18.461	0.951
	(50,5,5)	1.374	18.461	0.951
	(100,2,2)	1.368	18.430	0.951
	(100,5,2)	1.368	18.430	0.951
	(100,5,5)	1.368	18.430	0.951

Table 12: Estimates of percent bias, MSE, and coverage varying the percentages of missing values and number of imputations at each stage for one type of nonignorable missing value when there is missingness in three different variables

4.4.4 Sensitivity Analysis

A final simulation was performed to assess how sensitive the results are to the choice of k. The choice of k, the amount by which the ignorable imputed values are pertubed, is selected at the discretion of the researcher based on an assumption about the missing values (Siddique et al., 2012). The sensitivity of that assumption is tested using four different values for k.

Set-Up

The data for the simulations is generated through the same process as Section 4.4.1. The missingness structure is as follows:

1. The first type of missing value is created with a MNAR structure. Let $MNAR\%$ denote the percentage of missing values due to the type of missingness which is MNAR. Values in Y are removed if they are above the top $MNAR\%$ percentile of Y.

2. The second type of missing value is MAR. Values in Y are removed if they are below the bottom $MAR\%$ percentile of X_1.

3. The final type of missing value is MCAR. Values in Y are randomly removed so that the percentage of missing values due to this type of missingness is $MCAR\%$.

The percent breakdown for the missing values is the same as in Section 4.4.1 and can be found in Table 8. The missing values are imputed using the norm package in R and the order of imputation is $MNAR\%$, $MAR\%$, $MCAR\%$. The four values of k used are:

1. k=0.8 (indicating an assumption that the nonignorable missing values are 20% lower than their ignorable counterparts)

2. k=1.0 (indicating that the nonignorable imputations are the same as the ignorable counterparts, i.e. assuming MAR)

3. k=1.2 (indicating nonignorable imputations are 20% higher than the ignorable)

4. k=1.4 (indicating nonignorable imputations are 40% higher).

Based on the knowledge that the MNAR missing values are missing based on the upper percentile of Y, the expectation is that the nonignorable imputed values are higher than the ignorable counterparts.

Results

Simulation results are presented in four tables, one for each overall percentage of missing values. Table 13 is included here as an example and the remaining tables may be found in Appendix B. All of these tables display a clear pattern indicating the importance of a correct choice of k. When the percentage of missing values is as high as 75% (Table 13), misspecification of k in the incorrect direction (i.e. when k=0.8) yields percent bias as high as 14.85% with extremely high MSE (between 570 and 605) and extremely low coverage (between 1.1% and 0.2%) compared to the other choices of k. The choice of

k=1.2 yields the lowest percent bias, lowest MSE, and highest coverage for the 75% missingness rate.

The same results hold true for the other percentages of missingness (Tables 23, 24, and 25). The differences are less pronounced as the percentage of missingness decreases but the analysis is quite sensitive to the choice of k. Siddique et al. (2012) discuss treating the choice of k as missing information which can be imputed in a separate stage in the form of draws from some distribution. By choosing k in that way, the missing information due to the uncertainty of k is incorporated into the analysis. Such an approach was not examined here since the primary focus was simply to determine whether or not the choice of k had any impact on the analyses. Care must be taken on the part of the researcher to consider where the nonignorable missing values are expected to lie compared to the ignorable imputed values. Sensitivity analysis should be performed to assess the impact of k when applying this work to real data sets.

k	(L,M,N)	Percent Bias (%)	MSE	Coverage
k=0.8	(10,2,2)	-14.848	596.696	0.011
	(10,5,2)	-14.954	604.289	0.004
	(10,5,5)	-14.963	604.934	0.002
	(50,2,2)	-14.481	567.316	0.003
	(50,5,2)	-14.560	573.025	0.003
	(50,5,5)	-14.567	573.527	0.003
	(100,2,2)	-14.528	570.558	0.002
	(100,5,2)	-14.569	573.570	0.002
	(100,5,5)	-14.576	574.063	0.002
k=1.0	(10,2,2)	-7.511	180.440	0.410
	(10,5,2)	-7.570	182.387	0.350
	(10,5,5)	-7.578	182.616	0.345
	(50,2,2)	-7.093	162.975	0.393
	(50,5,2)	-7.145	164.715	0.383
	(50,5,5)	-7.149	164.834	0.381
	(100,2,2)	-7.097	162.793	0.389
	(100,5,2)	-7.129	163.976	0.374
	(100,5,5)	-7.135	164.189	0.372
k=1.2	(10,2,2)	-0.165	43.004	0.932
	(10,5,2)	-0.227	42.575	0.929
	(10,5,5)	-0.237	42.481	0.925
	(50,2,2)	0.328	41.238	0.933
	(50,5,2)	0.267	40.870	0.936
	(50,5,5)	0.262	40.833	0.935
	(100,2,2)	0.328	40.694	0.938
	(100,5,2)	0.285	40.590	0.935
	(100,5,5)	0.276	40.572	0.935
k=1.4	(10,2,2)	7.178	184.909	0.752
	(10,5,2)	7.091	180.771	0.711
	(10,5,5)	7.076	180.069	0.703
	(50,2,2)	7.767	205.271	0.675
	(50,5,2)	7.677	201.220	0.686
	(50,5,5)	7.670	200.922	0.687
	(100,2,2)	7.753	203.747	0.675
	(100,5,2)	7.688	201.082	0.672
	(100,5,5)	7.676	200.607	0.672

Table 13: Estimates of percent bias, MSE, and coverage varying k for 75% missing values

Chapter 5

Application of Method

Clinical studies tend to have large amounts of missing values. This is especially true of longitudinal studies which require subjects to commit an extended period of time to participation in the study. Moreover, the missing values are often of different types including subjects who are lost to follow-up, subjects who have intermittent missed follow-ups, failure to observe a covariate, measurement bias, or censored values. In the presence of so much missingness, using a principled missing data method is critical to the analysis of the data. In this chapter, the three-stage multiple imputation method is applied to a data set associated with a sleep study.

5.1 Sleep Study Design

Using data provided by a large pharmaceutical company, we analyzed the data from a clinical trial related to a Phase 2 study for the treatment of chronic insomnia. The study was a randomized double-blind parallel group placebo-controlled study with multiple doses of a compound used for treating insomnia. The study lasted for a period of four weeks and had an additional one week of placebo run-in before the randomization.

There were five treatment groups including placebo, 15, 30, 45, and 60 mg of the active compound. The total sample size for the study was 672.

The data collection was longitudinal with repeated measures and relied on patient-reported outcomes. Covariates of age, race, sex, and clinical site were also included. The collection of the data was handled using a data collection phone system where the subjects would phone each day and respond to five questions regarding the previous night's sleep.

The primary response of interest in this study was the subjective wake time after sleep onset (SWASO) which was measured in minutes. The SWASO measurements were derived from the daily phone responses and were averaged for each of the weeks of the study. If fewer than four days were recorded by the phone collection system, the SWASO for that week was coded as missing.

There are two main effects of interest: the immediate effect of the treatment and the persistent effect of the treatment. The immediate effect is classified as the difference between baseline and Week 1 for each dose and the persistent effect is classified as the difference between baseline and Week 4 for each dose.

Day	Percent Missing (%)	Day	Percent Missing (%)
1	9.8	15	18.0
2	12.1	16	21.3
3	13.5	17	23.1
4	12.9	18	24.4
5	11.9	19	24.0
6	11.6	20	22.0
7	12.8	21	23.5
8	11.2	22	21.0
9	18.3	23	25.3
10	18.6	24	26.6
11	18.9	25	28.0
12	19.6	26	28.0
13	18.0	27	26.3
14	18.2	28	28.9

Table 14: Proportion of missing daily SWASO values

5.2 Missing Data

Figures 3 and 4 display the missing data rates for the daily SWASO measurements for all the subjects and for the placebo and high dose group, respectively. Figure 4 illustrates that there was a higher rate of missing values for those on the high dose of the compound. The rate of missingness increased over time and the highest percentage of missing values for any study day was 28.9% on day 28. Table 14 displays the actual percentages of missing values for each study day.

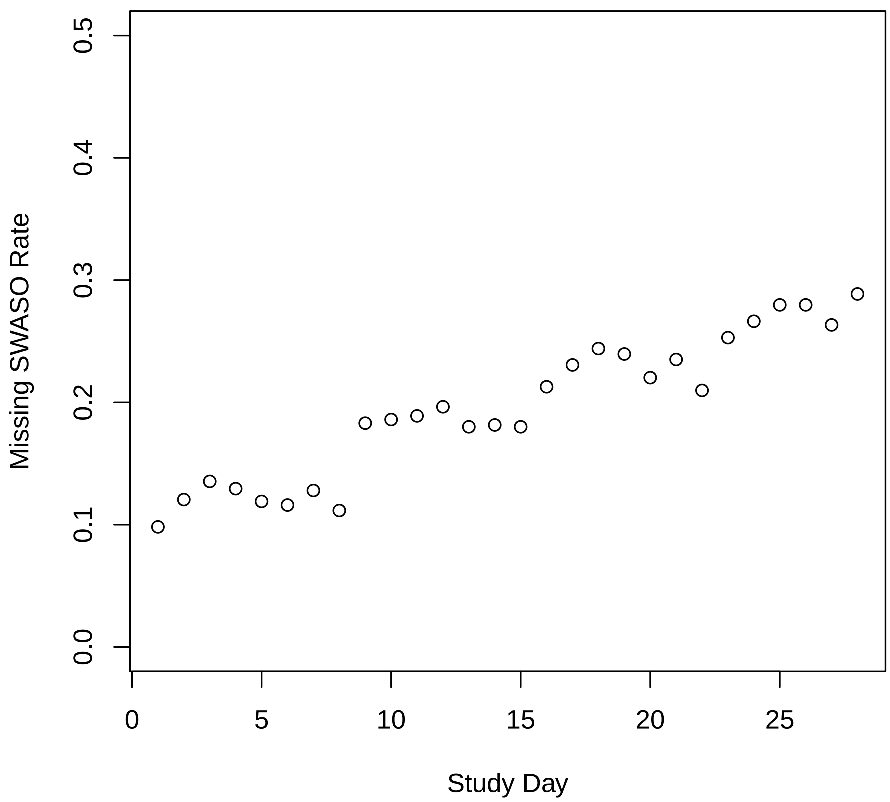

Figure 3: Proportion of missing daily SWASO measurements for study days 1-28 for all subjects

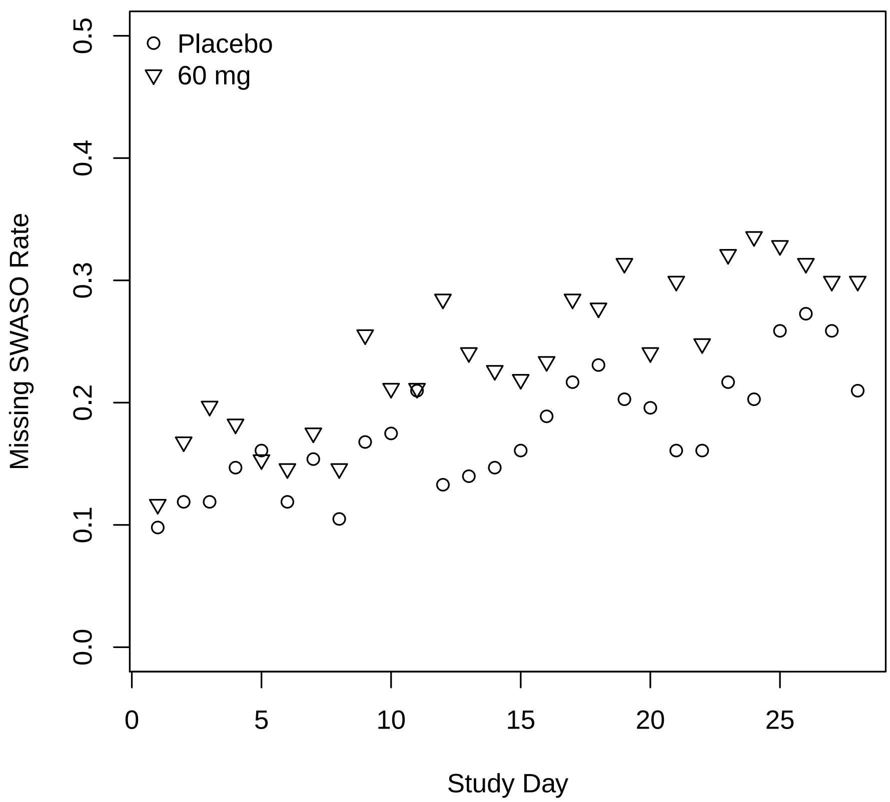

Figure 4: Proportion of missing daily SWASO measurements for study days 1-28 for placebo group and high dose group

	Dose				
	0	15	30	45	60
COMPLETED STUDY	124	110	115	104	109
ADVERSE EVENT	3	4	1	5	13
SUBJECT DIED	0	0	0	0	1
PROTOCOL VIOLATION	4	5	2	3	3
LOST TO FOLLOW UP	1	1	3	5	4
OTHER	2	0	2	0	1
FAILED ENTRANCE CRITERIA	1	0	1	0	0
SUBJECT WITHDREW CONSENT	8	13	10	7	6
PREGNANCY	0	1	0	0	0

Table 15: Reason for end of dosing

The weekly SWASO measurements were the average of the daily SWASO values for that week. A weekly SWASO was coded as missing if a subject provided fewer than four daily measurements for that week. Table 15 summarizes the reasons for dropping out of the study, broken down by dose group. There were 36 subjects who were missing all of the weekly SWASO measurements. Of those 36, 13 reported an adverse event. Figure 5 provides a summary of the missing data patterns for the weekly SWASO measurements where grey regions represent missing data and white regions represent observed data. The figure was created using the R package VIM (Templ et al., 2013). The most common missing data patterns were those corresponding to monotone dropout.

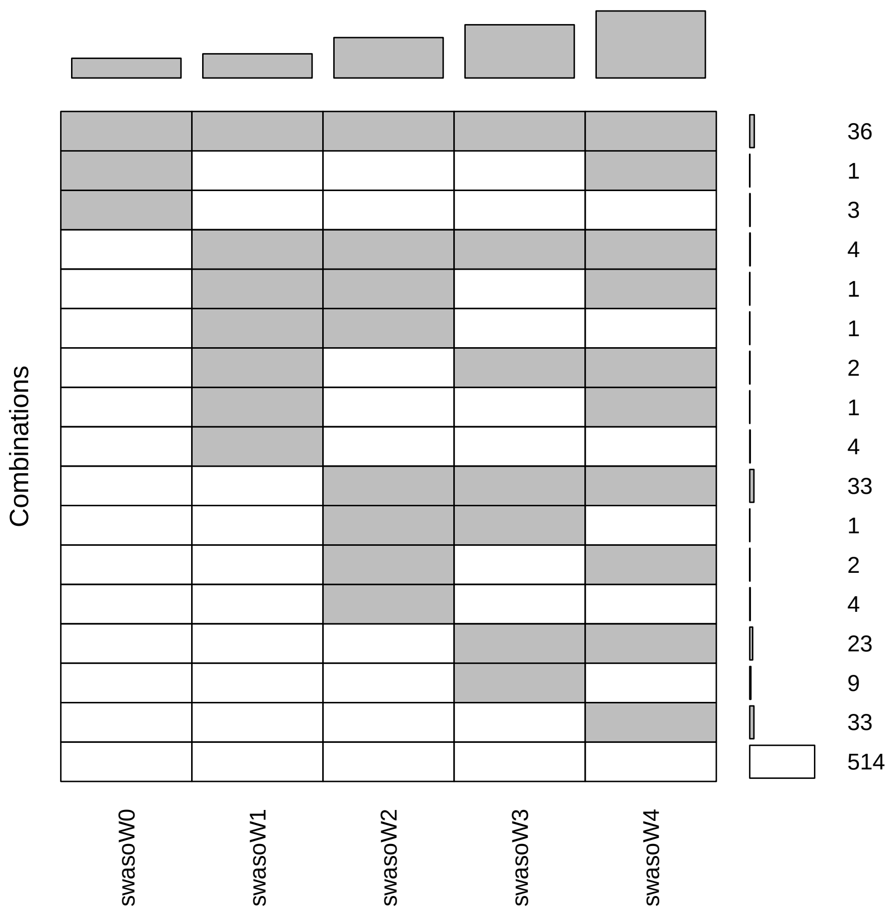

Figure 5: Missing data patterns for the weekly averaged SWASO measurements. The grey regions represent the missing data and the numbers on the right represent the frequencies of each pattern

5.3 Methods

A common method of dealing with missing data in sleep studies is last observation carried forward (LOCF). Despite being widely discredited (Olsen et al., 2012), it remains a popular approach in handling incomplete data. Due to its popularity, LOCF was included in the comparison of different methods for analyzing the data. The four methods used were complete case analysis (CC), last observation carried forward (LOCF), standard multiple imputation (MI), and three-stage multiple imputation (3MI). All of these methods were discussed in Sections 1.1.2 and 2.1.

There are two main responses of interest. The first is the immediate effect model in which the response of interest is the Week 1 change from baseline. The second is the persistent effect model in which the response of interest is the Week 4 change from baseline. The analysis model for the immediate effect is

$$\text{Week 1 - Baseline} = \text{Baseline} + \text{Treatment} + \text{Center} \tag{5.1}$$

where the treatment variable is a categorical indicator variable. There are five treatment groups so there are four β parameters corresponding to the treatment variable with placebo used as the reference group. A treatment is considered significant if at least one of the five treatment groups is significant. The analysis model for the immediate effect expressed with all of the parameters is

$$
\begin{aligned}
\text{Week 1 - Baseline} =& \beta_0 + \beta_1(\text{Baseline}) + \beta_2(\text{Treatment15}) + \beta_3(\text{Treatment30}) \\
&+ \beta_4(\text{Treatment45}) + \beta_5(\text{Treatment60}) + \boldsymbol{\beta}(\text{Center}) + \epsilon
\end{aligned} \tag{5.2}
$$

where $\boldsymbol{\beta}$ is a vector of β parameters corresponding to each center and $\epsilon \overset{iid}{\sim} N(0, \sigma^2 I)$.

The analysis model for the fixed effect is

$$
\text{Week 4 - Baseline} = \text{Baseline} + \text{Treatment} + \text{Center} \tag{5.3}
$$

where the variables are as previously described. The analysis model for the persistent effect expressed with all of the parameters is

$$
\begin{aligned}
\text{Week 4 - Baseline} =& \beta_0 + \beta_1(\text{Baseline}) + \beta_2(\text{Treatment15}) + \beta_3(\text{Treatment30}) \\
&+ \beta_4(\text{Treatment45}) + \beta_5(\text{Treatment60}) + \boldsymbol{\beta}(\text{Center}) + \epsilon
\end{aligned} \tag{5.4}
$$

where $\boldsymbol{\beta}$ is a vector of β parameters corresponding to each center and $\epsilon \overset{iid}{\sim} N(0, \sigma^2 I)$.

The four methods compared the β estimates, standard errors, confidence intervals, widths of confidence intervals, and p-values for each of the two models.

Table 16 displays the number of subjects from each treatment group included in each of the four analyses.

Table 16: Number of subjects included in analysis by treatment for each of the four methods

Method	Model	Placebo	15 mg	30 mg	45 mg	60 mg
CC	Immediate	132	126	127	116	119
	Persistent	119	103	115	106	108
LOCF	Immediate	134	128	128	118	124
	Persistent	134	128	128	118	124
MI	Immediate	143	134	134	124	137
	Persistent	143	134	134	124	137
3MI	Immediate	143	134	134	124	137
	Persistent	143	134	134	124	137

5.3.1 Complete Case Analysis

There were 672 subjects enrolled and randomized in the study. Of those 672 subjects, 40 were missing baseline SWASO values, 48 were missing Week 1 SWASO values, and 117 were missing Week 4 SWASO values (based on the fewer than 4 observations rule). For the immediate effect model, there were 620 subjects with complete data for baseline and Week 1. For the persistent effect model, there were 551 subjects with complete data for baseline and Week 4. For the immediate effect model, 7.7% of the data were missing and for the persistent effect model, 18.0% of the data were missing.

5.3.2 Last Observation Carried Forward

LOCF was implemented as described in Section 1.1.2 in which missing values were replaced with the last observed values. In the case of missing baseline values, if there were fewer than 4 observations in the baseline week, baseline was coded as missing. If baseline was missing, those subjects were removed from the analysis. There were 40 such subjects that were removed leaving a total of 632 subjects for the analysis.

5.3.3 Standard Multiple Imputation

Daily SWASO values were imputed using the R package pan (Schafer, 2013). Pan is a package that allows for imputation on multivariate panel data or clustered data. It is used for data which has multiple variables collected on individuals over time (Harel & Zhou, 2007). The imputation model used center, age, sex, finishing status, race, and dose as predictors for the imputations. Missing values were assumed to be ignorable. There were 4024 missing values out of a total 23520 possible daily SWASO values so the percentage of missing values was 17.11%. A total of 40 imputations were used in order for there to be consistency with the number of imputations to be used in three-stage multiple imputation. Daily SWASO values were imputed and then averaged to form the weekly SWASO values for use in the analysis models.

5.3.4 Three-Stage Multiple Imputation

The missing daily SWASO values can be partitioned into three types. The first type of missing value corresponds to subjects who dropped out of the study. These subjects failed to complete the study for a variety of reasons (summarized in Table 15) but were treated as one group of missing values. These missing values were assumed to be nonignorable. Since several subjects experienced adverse events or dropped out due to lack of efficacy, the assumption used for the nonignorable imputations was that the

Table 17: Number and percentages of missing daily SWASO values by type of missing value

Observed SWASO Values	Dropouts (First Type)	Intermittent Missed (Second Type)	Baseline Missed (Third Type)
19496 82.89%	2012 8.55%	1440 6.12%	572 2.43%

nonignorable values were 20% higher (with higher corresponding to more time awake) than their ignorable counterparts. The assumption of k=1.2 was addressed in Section 5.5 which performed a sensitivity analysis to assess the impact of that assumption. The second type of missing value corresponds to subjects who completed the study but had some intermittent missing values. These missing values were assumed to be MAR and ignorable. The third type of missing value corresponds to subjects who were missing baseline values. Baseline was used as a covariate in the model and these missing values were assumed to be MAR and ignorable. Table 17 summarizes how many missing values are of each type.

Missing values were imputed using the R package pan (Schafer, 2013) with dropouts being imputed first, intermittent missed observations being imputed second, and baseline values being imputed last. The nonignorable dropouts must be imputed first because the data were $CMAR^2$ as described in Definition 4.8. The predictor variables used in the imputation model were center, age, sex, race, finishing status, and dose. The number of imputations used were $L = 10$, $M = 2$, and $N = 2$ for a total of 40 completed data sets. These numbers were chosen based on the simulation results in Chapter 4. Daily SWASO values were imputed and then averaged into weekly SWASO values for use in the analysis models.

5.4 Results

Both the immediate effect model and the persistent effect model showed differing results between methods. Table 18 and Table 19 provide estimates for the baseline coefficient and each of the treatment coefficients, along with standard errors, 95% confidence intervals, and widths of those confidence intervals. CC and LOCF showed that baseline and all treatments except for 15 mg were significant. However, MI and 3MI only found baseline to be significant and no significant treatment effects. The standard errors were lower and the confidence intervals were narrower for MI and 3MI compared to the other two methods.

It is unsurprising that CC and LOCF would show treatment effects where MI and 3MI found none due to the fact that many of the missing values are due to lack of efficacy of the treatment. One would expect the SWASO values to be higher among the people who drop out which biases the results toward a treatment effect (i.e. subjects who stay in the study may be continuing because the treatment is efficacious). Differing results such as these call into question analysis methods that use only observed values or use the outdated method of LOCF.

One of the benefits of three-stage multiple imputation is the ability to quantify the rates of missing information by each type of missing value. Table 20 summarizes those

		Parameter Estimate	Standard Error	Lower Bound	Upper Bound	Width
CC	Baseline	-0.359*	0.032	-0.421	-0.297	0.124
	15 mg	-1.342	5.016	-11.195	8.510	19.705
	30 mg	-13.263*	5.031	-23.144	-3.381	19.763
	45 mg	-13.912*	5.172	-24.073	-3.752	20.321
	60 mg	-25.825*	5.115	-35.873	-15.777	20.096
LOCF	Baseline	-0.350*	0.031	-0.411	-0.288	0.122
	15 mg	-0.452	4.961	-10.197	9.293	19.491
	30 mg	-13.175*	4.987	-22.970	-3.380	19.589
	45 mg	-12.376*	5.107	-22.406	-2.345	20.061
	60 mg	-24.964*	5.023	-34.831	-15.097	19.735
MI	Baseline	-1.082*	0.039	-1.158	-1.005	0.153
	15 mg	1.295	2.882	-4.355	6.946	11.302
	30 mg	2.848	2.878	-2.794	8.490	11.284
	45 mg	-4.489	2.932	-10.237	1.260	11.497
	60 mg	1.711	2.900	-3.976	7.398	11.374
3MI	Baseline	-1.073*	0.036	-1.144	-1.002	0.142
	15 mg	1.593	2.918	-4.129	7.315	11.444
	30 mg	3.135	2.920	-2.590	8.861	11.452
	45 mg	-4.758	2.895	-10.433	0.917	11.350
	60 mg	1.487	2.918	-4.235	7.209	11.444

Table 18: Immediate effect model results for each of the four methods used. Included are parameter estimates, standard errors, the lower and upper bounds of the 95% confidence intervals, and the width of the confidence intervals. * indicates significance at the 95% confidence level

rates of missing information. It is worthwhile to note that the largest overall rate of missing information was 15.11% and that for some of the types, the rate of missing information was practically zero. For the immediate effect model, most of the missing information was typically attributable to the first type of missing value (the dropout) but for the persistent effect model, most of the missing information was attributable to the second type of missing value (the intermittent missed values among completers). The missing information due to the baseline value is negligible except in the case of the estimate of the baseline coefficient. The implication here is that, when the immediate effect is of the most interest, care should be taken to retain subjects who would have dropped out. Otherwise, it is valuable to seek to minimize the number of intermittent missed values.

		Parameter Estimate	Standard Error	Lower Bound	Upper Bound	Width
CC	Baseline	-0.494*	0.039	-0.570	-0.417	0.152
	15 mg	-2.693	6.253	-14.980	9.593	24.572
	30 mg	-20.671*	6.110	-32.676	-8.666	24.011
	45 mg	-18.476*	6.280	-30.817	-6.136	24.681
	60 mg	-19.577*	6.200	-31.760	-7.394	24.366
LOCF	Baseline	-0.463*	0.037	-0.535	-0.391	0.144
	15 mg	0.148	5.845	-11.334	11.629	22.963
	30 mg	-16.460*	5.875	-28.000	-4.920	23.079
	45 mg	-18.643*	6.016	-30.460	-6.825	23.635
	60 mg	-21.188*	5.918	-32.814	-9.563	23.250
MI	Baseline	-0.991*	0.037	-1.063	-0.918	0.145
	15 mg	-1.056	2.794	-6.535	4.422	10.958
	30 mg	-0.892	2.954	-6.689	4.905	11.593
	45 mg	1.012	2.941	-4.757	6.780	11.537
	60 mg	-0.774	2.838	-6.341	4.793	11.133
3MI	Baseline	-0.995*	0.037	-1.067	-0.922	0.145
	15 mg	-0.973	2.726	-6.317	4.370	10.687
	30 mg	-0.175	2.785	-5.636	5.286	10.921
	45 mg	1.007	3.029	-4.945	6.959	11.904
	60 mg	-0.745	2.726	-6.087	4.598	10.685

Table 19: Persistent effect model results for each of the four methods used. Included are parameter estimates, standard errors, the lower and upper bounds of the 95% confidence intervals, and the width of the confidence intervals. * indicates significance at the 95% confidence level

		$\hat{\lambda}$ (%)	$\hat{\lambda}^{A}$ (%)	$\hat{\lambda}^{B\|A}$ (%)	$\hat{\lambda}^{C\|A,B}$ (%)
Immediate	Baseline	3.411	2.766	0.000	1.192
	15 mg	6.145	5.527	0.613	0.005
	30 mg	5.663	5.198	0.465	0.000
	45 mg	0.503	0.159	0.328	0.016
	60 mg	7.114	4.291	2.819	0.005
Persistent	Baseline	14.659	0.000	13.941	8.373
	15 mg	1.283	0.000	1.723	0.044
	30 mg	4.438	3.363	1.075	0.000
	45 mg	15.113	9.036	6.074	0.003
	60 mg	2.279	0.000	4.322	0.016

Table 20: Estimated rates of missing information by type of missing value

5.5 Sensitivity Analysis

Nonignorable imputations used in 3MI were perturbed by k=1.2, under the assumption that the nonignorable imputed values would be 20% higher than the ignorable counterparts. This assumption is one made by the researcher based on knowledge of the nature of the data. This section performs a sensitivity analysis to examine if varying values for k significantly impacted the results.

Tables 21 and 22 display the results for the immediate effect model and the persistent effect model with different values of k. The values of k used were k=0.8 (the nonignorable values are 20% lower), k=1.0 (the nonignorable values are the same as the ignorable values), k=1.2 (the nonignorable values are 20% higher), and k=1.4 (the nonignorable values are 40% higher). The analyses showed that even when k is specified to be 0.8 when the nonignorable values were believed to be higher, there was no impact on the significance of the parameters. As k increased, the confidence intervals got wider but there was otherwise no trend in changing k. The choice of k did not have a large impact on the analysis of this data.

		Parameter Estimate	Standard Error	Lower Bound	Upper Bound	Width
	Baseline	-1.073*	0.034	-1.139	-1.006	0.134
	15 mg	1.441	2.747	-3.943	6.826	10.769
k=0.8	30 mg	3.123	2.744	-2.257	8.502	10.759
	45 mg	-4.365	2.752	-9.758	1.028	10.786
	60 mg	1.965	2.755	-3.436	7.366	10.802
	Baseline	-1.072*	0.035	-1.141	-1.004	0.137
	15 mg	1.519	2.823	-4.016	7.055	11.071
k=1.0	30 mg	3.128	2.822	-2.405	8.661	11.066
	45 mg	-4.563	2.813	-10.076	0.951	11.028
	60 mg	1.727	2.827	-3.817	7.270	11.087
	Baseline	-1.073*	0.036	-1.144	-1.002	0.142
	15 mg	1.593	2.918	-4.129	7.315	11.444
k=1.2	30 mg	3.135	2.920	-2.590	8.861	11.452
	45 mg	-4.758	2.895	-10.433	0.917	11.350
	60 mg	1.487	2.918	-4.235	7.209	11.444
	Baseline	-1.074*	0.038	-1.148	-1.000	0.148
	15 mg	1.664	3.028	-4.276	7.604	11.880
k=1.4	30 mg	3.146	3.036	-2.808	9.100	11.908
	45 mg	-4.951	2.996	-10.824	0.922	11.746
	60 mg	1.247	3.025	-4.686	7.180	11.866

Table 21: Immediate effect model results for 3MI varying k. Included are parameter estimates, standard errors, the lower and upper bounds of the 95% confidence intervals, and the width of the confidence intervals. * indicates significance at the 95% confidence level

		Parameter Estimate	Standard Error	Lower Bound	Upper Bound	Width
	Baseline	-0.995*	0.034	-1.063	-0.927	0.135
	15 mg	-1.060	2.537	-6.031	3.912	9.943
k=0.8	30 mg	-0.407	2.585	-5.474	4.661	10.135
	45 mg	1.400	2.788	-4.073	6.874	10.947
	60 mg	-0.918	2.548	-5.912	4.076	9.988
	Baseline	-0.995*	0.035	-1.064	-0.925	0.139
	15 mg	-1.016	2.613	-6.138	4.105	10.243
k=1.0	30 mg	-0.291	2.668	-5.522	4.939	10.462
	45 mg	1.204	2.890	-4.473	6.881	11.354
	60 mg	-0.832	2.619	-5.966	4.302	10.268
	Baseline	-0.995*	0.037	-1.067	-0.922	0.145
	15 mg	-0.973	2.726	-6.317	4.370	10.687
k=1.2	30 mg	-0.175	2.785	-5.636	5.286	10.921
	45 mg	1.007	3.029	-4.945	6.959	11.904
	60 mg	-0.745	2.726	-6.087	4.598	10.685
	Baseline	-0.995*	0.039	-1.071	-0.918	0.153
	15 mg	-0.930	2.872	-6.559	4.698	11.257
k=1.4	30 mg	-0.058	2.932	-5.807	5.691	11.498
	45 mg	0.810	3.201	-5.481	7.100	12.580
	60 mg	-0.657	2.862	-6.268	4.954	11.221

Table 22: Persistent effect model results for 3MI varying k. Included are parameter estimates, standard errors, the lower and upper bounds of the 95% confidence intervals, and the width of the confidence intervals. * indicates significance at the 95% confidence level

Chapter 6

Conclusion and Future Work

Overview

The problem of incomplete data is one which researchers must handle on a regular basis. Many researchers fail to consider missing values of varying natures in their analyses, treating them as a singular type or not considering the impact of the missing values at all. Studies tend to record information describing different types of nonresponse but rarely use that information in the analysis. This dissertation builds the statistical methodology to incorporate analyzing data with values missing for three distinct reasons.

The main contributions of this dissertation are (a) an extension of the multiple imputation methodology of Rubin (1976) and Shen (2000); (b) the derivation of the combining rules for three-stage multiple imputation; (c) the derivation of the estimates and asymptotic distributions for the rates of missing information with three types of missing values; (d) an extension of ignorability conditions for missing values of three different types; and (e) an application of the methodology to a sleep study with comparisons to other commonly used missing data methods.

Future Work

The work of this dissertation can be extended and applied in many different ways. The first, and most obvious, extension is considering four or more types of missing values. The derivation of the combining rules for k-stage multiple imputation is straightforward and the results may be found in Appendix A and Equations (A.1) and (A.2). A discussion of the merit of allowing for k stages of multiple imputation would be useful in determining when the model is being over-specified and when the complexity of interpretation outweighs the benefits of many stages. While there may be some benefits to k-stage multiple imputation, the number of imputations required and the computation time would increase dramatically with additional stages.

A second interesting area of extension is allowing for the missing values to be of different forms. One could classify one type of missing value as a missing latent class, model uncertainty, or uncertainty in the value of k for nonignorable imputations in the style of Siddique et al. (2012). One type of missing value may be random draws from a distribution of k which is prespecified. This would account for missing information contributed by the uncertainty of k and could reduce the strength of the assumption being made.

Another area that may be explored is the incorporation of pattern-mixture models (Hedeker & Gibbons, 1997; Demirtas & Schafer, 2003). Given the extensions to the ignorability conditions presented in Chapter 4, a logical extension would be to model the missingness process with three types of missing values simultaneously using a pattern-mixture model.

The derivations and analyses presented in this dissertation are valid under the large data assumption. In standard multiple imputation, there exist small-sample degrees of freedom (Barnard & Rubin, 1999; Reiter & Raghunathan, 2007). Small-sample degrees of freedom have not yet been explored for either two-stage multiple imputation or three-stage multiple imputation. Development of those small-sample combining rules would increase the areas of applicability of these methods.

Finally, multiple imputation in three stages can be implemented in different types of studies. Chapter 5 addressed a longitudinal sleep study but it would be valuable to apply this work to other types of studies including survey studies and complex, large data sets.

All of these areas of future research are worthwhile extensions of the work presented in this dissertation which provides a foundation for the expansion of these ideas in the statistical methodology. How to appropriately handle incomplete data with three types of missing values is an applicable and fundamental contribution to the missing data literature.

Appendix A

Combining Rules for k-Stage Multiple Imputation

The proof for the combining rules for k-stage multiple imputation is a straightforward extension of the proof for three-stage multiple imputation. The details of the proof are not included here but directly follow the steps presented in Chapter 2 with additional variables. The process for creating the imputations would be similar but would have additional nested steps for creating the completed data sets. A discussion of the benefits and complications of including k types of missing values is beyond the scope of this dissertation but the results for the combining rules are presented here.

Let $N_1, N_2, \ldots, N_k$ denote the number of imputations at each of the stages of imputation for a total of $N_1 \times N_2 \times \ldots \times N_k$ completed data sets. Let $\bar{Q}$ be the average of the parameter estimates from each of the completed data sets and $\bar{U}$ be the average of the complete data variance estimates. Let MS_i denote the mean square from the i^{th} nest in the nested ANOVA decomposition where $i = 1, \ldots, k$. Then, the total variance, T, is

$$T = \bar{U} + \frac{1}{\prod_{j=2}^{k} N_j} \left(1 + \frac{1}{N_1}\right) MS_1 + \sum_{i=2}^{k-1} \frac{1}{\prod_{j=i+1}^{k} N_j} \left(1 - \frac{1}{N_i}\right) MS_i + \left(1 - \frac{1}{N_k}\right) MS_k \tag{A.1}$$

with degrees of freedom

$$\begin{aligned}
\nu^{-1} = & \left[\frac{\frac{1}{\prod_{j=2}^{k} N_j}\left(1 + \frac{1}{N_1}\right) MS_1}{T}\right]^2 [(N_1 - 1)]^{-1} \\
& + \sum_{i=2}^{k-1} \left[\left[\frac{\frac{1}{\prod_{j=i+1}^{k} N_j}\left(1 - \frac{1}{N_i}\right) MS_i}{T}\right]^2 \left[\left(\prod_{j=1}^{k-2} N_j\right)(N_i - 1)\right]^{-1}\right] \\
& + \left[\frac{\left(1 - \frac{1}{N_k}\right) MS_k}{T}\right]^2 \left[\left(\prod_{j=1}^{k-1} N_j\right)(N_k - 1)\right]^{-1}.
\end{aligned} \tag{A.2}$$

These formulae hold for three or more types of missing values. When $k = 3$, these formulae reduce to the three-stage multiple imputation combining rules in Equations (2.45) and (2.46).

Appendix B

Supplemental Tables

The tables in this appendix are supplements to the sensitivity analysis presented in Section 4.4.4. Each table corresponds to a different percentage of missing values (50%, 40%, and 15%, respectively).

k	(L,M,N)	Percent Bias (%)	MSE	Coverage
	(10,2,2)	-10.084	278.516	0.023
	(10,5,2)	-10.131	280.923	0.019
	(10,5,5)	-10.102	279.418	0.017
	(50,2,2)	-10.100	279.229	0.014
k=0.8	(50,5,2)	-10.065	277.492	0.017
	(50,5,5)	-10.033	275.841	0.018
	(100,2,2)	-10.140	281.403	0.016
	(100,5,2)	-10.126	280.707	0.015
	(100,5,5)	-10.110	279.892	0.015
	(10,2,2)	-5.674	101.585	0.362
	(10,5,2)	-5.740	103.418	0.338
	(10,5,5)	-5.713	102.612	0.342
	(50,2,2)	-5.707	102.303	0.319
k=1.0	(50,5,2)	-5.684	101.689	0.325
	(50,5,5)	-5.657	100.895	0.326
	(100,2,2)	-5.800	105.146	0.313
	(100,5,2)	-5.790	104.876	0.311
	(100,5,5)	-5.777	104.488	0.314
	(10,2,2)	-1.270	25.016	0.935
	(10,5,2)	-1.364	25.403	0.930
	(10,5,5)	-1.330	25.168	0.932
	(50,2,2)	-1.339	25.003	0.925
k=1.2	(50,5,2)	-1.316	24.850	0.925
	(50,5,5)	-1.284	24.639	0.926
	(100,2,2)	-1.468	26.050	0.919
	(100,5,2)	-1.458	25.980	0.916
	(100,5,5)	-1.443	25.869	0.916
	(10,2,2)	3.129	48.455	0.930
	(10,5,2)	2.997	45.857	0.926
	(10,5,5)	3.044	46.584	0.927
	(50,2,2)	3.018	45.901	0.919
k=1.4	(50,5,2)	3.048	46.281	0.920
	(50,5,5)	3.089	46.927	0.918
	(100,2,2)	2.863	43.707	0.921
	(100,5,2)	2.873	43.834	0.922
	(100,5,5)	2.893	44.118	0.919

Table 23: Estimates of percent bias, MSE, and coverage varying k for 50% missing values

k	(L,M,N)	Percent Bias (%)	MSE	Coverage
k=0.8	(10,2,2)	-9.970	270.158	0.006
	(10,5,2)	-10.115	277.588	0.005
	(10,5,5)	-10.106	277.163	0.005
	(50,2,2)	-10.468	296.163	0.005
	(50,5,2)	-10.442	294.772	0.005
	(50,5,5)	-10.429	294.057	0.005
	(100,2,2)	-10.422	293.664	0.005
	(100,5,2)	-10.417	293.413	0.005
	(100,5,5)	-10.426	293.857	0.005
k=1.0	(10,2,2)	-5.431	92.572	0.301
	(10,5,2)	-5.595	97.160	0.296
	(10,5,5)	-5.583	96.844	0.304
	(50,2,2)	-6.036	110.201	0.232
	(50,5,2)	-6.008	109.343	0.239
	(50,5,5)	-5.999	109.062	0.239
	(100,2,2)	-5.991	108.779	0.242
	(100,5,2)	-5.980	108.467	0.244
	(100,5,5)	-5.990	108.759	0.243
k=1.2	(10,2,2)	-0.876	20.657	0.923
	(10,5,2)	-1.083	21.619	0.930
	(10,5,5)	-1.067	21.551	0.932
	(50,2,2)	-1.613	25.150	0.896
	(50,5,2)	-1.579	24.866	0.900
	(50,5,5)	-1.570	24.796	0.899
	(100,2,2)	-1.560	24.672	0.903
	(100,5,2)	-1.545	24.562	0.903
	(100,5,5)	-1.557	24.657	0.903
k=1.4	(10,2,2)	3.698	55.599	0.851
	(10,5,2)	3.426	50.493	0.877
	(10,5,5)	3.447	50.887	0.879
	(50,2,2)	2.803	40.383	0.906
	(50,5,2)	2.848	41.019	0.902
	(50,5,5)	2.858	41.165	0.900
	(100,2,2)	2.870	41.291	0.903
	(100,5,2)	2.891	41.608	0.901
	(100,5,5)	2.875	41.370	0.900

Table 24: Estimates of percent bias, MSE, and coverage varying k for 40% missing values

k	(L,M,N)	Percent Bias (%)	MSE	Coverage
	(10,2,2)	-2.786	34.031	0.716
	(10,5,2)	-2.832	34.747	0.712
	(10,5,5)	-2.868	35.276	0.711
	(50,2,2)	-2.641	32.050	0.743
k=0.8	(50,5,2)	-2.639	32.028	0.742
	(50,5,5)	-2.633	31.948	0.742
	(100,2,2)	-2.633	31.918	0.745
	(100,5,2)	-2.630	31.893	0.741
	(100,5,5)	-2.631	31.899	0.741
	(10,2,2)	-1.777	22.527	0.825
	(10,5,2)	-1.823	23.002	0.824
	(10,5,5)	-1.858	23.336	0.822
	(50,2,2)	-1.595	20.999	0.853
k=1.0	(50,5,2)	-1.593	20.989	0.853
	(50,5,5)	-1.587	20.943	0.853
	(100,2,2)	-1.590	20.928	0.856
	(100,5,2)	-1.587	20.913	0.856
	(100,5,5)	-1.587	20.916	0.856
	(10,2,2)	-0.768	16.259	0.928
	(10,5,2)	-0.817	16.507	0.931
	(10,5,5)	-0.854	16.670	0.930
	(50,2,2)	-0.550	15.571	0.938
k=1.2	(50,5,2)	-0.548	15.571	0.939
	(50,5,5)	-0.542	15.555	0.939
	(100,2,2)	-0.546	15.525	0.940
	(100,5,2)	-0.543	15.525	0.939
	(100,5,5)	-0.543	15.526	0.939
	(10,2,2)	0.239	15.209	0.956
	(10,5,2)	0.185	15.206	0.959
	(10,5,5)	0.144	15.174	0.960
	(50,2,2)	0.494	15.744	0.959
k=1.4	(50,5,2)	0.496	15.757	0.959
	(50,5,5)	0.502	15.769	0.959
	(100,2,2)	0.497	15.709	0.961
	(100,5,2)	0.501	15.728	0.960
	(100,5,5)	0.500	15.727	0.960

Table 25: Estimates of percent bias, MSE, and coverage varying k for 15% missing values

Bibliography

BARNARD, J. (1995). *Cross-Match Procedures for Multiple-Imputation Inference: Bayesian Theory and Frequentist Evaluation.* Ph.D. thesis, Department of Statistics, University of Chicago, Chicago, IL.

BARNARD, J. & RUBIN, D. B. (1999). Small-sample degrees of freedom with multiple imputation. *Biometrika* **86,** 948–955.

BELIN, T. (2009). Missing data: What a little can do and what researchers can do in response. *American Journal of Opthalmology* **148,** 820–822.

DEMIRTAS, H. & SCHAFER, J. L. (2003). On the performance of random-coefficient pattern-mixture models for non-ignorable drop-out. *Statistics in Medicine* **22,** 2553–2575.

DEMPSTER, A., LAIRD, A. & RUBIN, D. B. (1997). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B (Methodological)* **39,** 1–38.

GELMAN, A., CARLIN, J., STERN, H. & RUBIN, D. B. (2003). *Bayesian Data Analysis.* Boca Raton, FL: Chapman and Hall.

HAREL, O. (2007). Inferences on missing information under multiple imputation and two-stage multiple imputation. *Statistical Methodology* **4,** 75–89.

HAREL, O. (2009). *Strategies for Data Analysis with Two Types of Missing Values: From Theory to Application.* Saarbrúcken, Germany: Lambert Academic Publishing.

HAREL, O., PELLOWSKI, J. & KALICHMAN, S. (2012). Are we missing the importance of missing values in HIV prevention randomized clinical trials? Review and recommendations. *AIDS and Behavior* **16,** 1382–1393.

HAREL, O. & SCHAFER, J. L. (2009). Partial and latent ignorability in missing-data problems. *Biometrika* **96,** 37–50.

HAREL, O. & ZHOU, X.-H. (2007). Multiple imputation: Review of theory, implementation, and software. *Statistics in Medicine* **26,** 3057–3077.

HEDEKER, D. & GIBBONS, R. (1997). Application of random-effects pattern-mixture models for missing data in longitudinal studies. *Psychological Methods* **2,** 64–78.

Lipsitz, S. R., Parzen, M. & Zhao, L. P. (2002). A degrees-of-freedom approximation in multiple imputation. *Journal of Statistical Computation and Simulation* **72,** 309–318.

Little, R. J. & Rubin, D. B. (2002). *Statistical Analysis with Missing Data.* New York: Wiley.

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Sciences* **9,** 538–573.

Molnar, F. J., Hutton, B. & Fergusson, D. (2008). Does analysis using "last observation carried forward" introduce bias in dementia research? *Canadian Medical Association Journal* **179,** 751–753.

Olsen, M. K., Stechuchak, K. M., Edinger, J. D., Ulmer, C. S. & Woolson, R. F. (2012). Move over LOCF: Principled methods for handling missing data in sleep disorder trials. *Sleep Medicine* **13,** 123–132.

Reiter, J. P. (2007). Small-sample degrees of freedom for multi-component significance tests with multiple imputation for missing data. *Biometrika* **94,** 502–508.

Reiter, J. P. & Raghunathan, T. E. (2007). The multiple adaptations of multiple imputation. *Journal of the American Statistical Association* **102,** 1462–1471.

Rubin, D. B. (1976). Inference and missing data. *Biometrika* **64,** 581–592.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys.* Hoboken, New Jersey: Wiley.

Schafer, J. & Novo, A. A. (2013). *Package 'norm': Analysis of multivariate normal datasets with missing values.* R Foundation for Statistical Computing. URL `http://cran.r-project.org/web/packages/norm/norm.pdf`.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data.* London: Chapman & Hall.

Schafer, J. L. (1999). Multiple imputation: A primer. *Statistical Methods in Medical Research* **8,** 3–15.

Schafer, J. L. (2013). *Multiple imputation for multivariate panel or clustered data.* R Foundation for Statistical Computing. URL `http://cran.r-project.org/web/packages/pan/pan.pdf`.

Schafer, J. L. & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods* **7,** 147–177.

Shen, Z. (2000). *Nested Multiple Imputation*. Ph.D. thesis, Department of Statistics, Harvard University, Cambridge, MA.

Siddique, J., Harel, O. & Crespi, C. (2012). Addressing missing data mechanism uncertainty using multiple-model multiple imputation: Application to a longitudinal clinical trial. *Annals of Applied Statistics* **6,** 1814–1837.

Templ, M., Alfons, A., Kowarik, A. & Prantner, B. (2013). *Visualization and imputation of missing values*. R Foundation for Statistical Computing. URL `http://cran.r-project.org/web/packages/VIM/VIM.pdf`.

Wagstaff, D. A. & Harel, O. (2011). A closer examination of three small-sample approximations to the multiple-imputation degrees of freedom. *The Stata Journal* **11,** 403–419.

White, I. & Carlin, J. (2010). Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Statistics in Medicine* **28,** 2920–2931.